(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平8－77294

(43)公開日　平成８年(1996)３月22日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｋ	9/46		Ｇ　9061－5H		
	9/20	３４０	Ｌ		
	9/62	６１０	Ｚ　9061－5H		

審査請求　未請求　請求項の数10　ＯＬ　（全 30 頁）

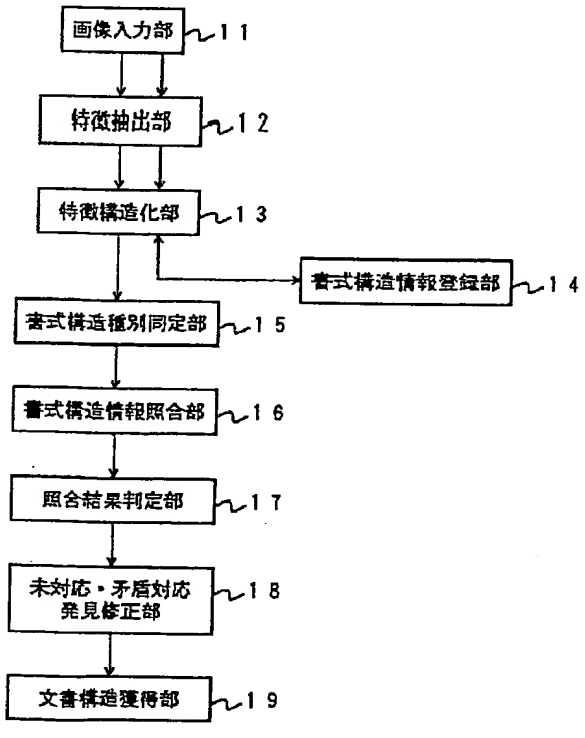
(21)出願番号　特願平6－212951

(22)出願日　平成６年(1994)９月６日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　石谷　康人
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内
(74)代理人　弁理士　鈴江　武彦

(54)【発明の名称】　文書画像処理装置

(57)【要約】
【目的】本発明は、帳票などの文書フォーマットを正確に特定でき、効率の良い文字列の抽出、読み取りを可能にした文書画像処理装置を提供する。
【構成】画像入力部１１より生成される帳票の入力画像に対して、特徴抽出部１２より抽出された図形特徴量を特徴構造化部でグループ化し、それぞれの特徴間の関係を抽出・管理する。構造化特徴と、書式構造種別同定部１５で予め登録されている処理対象文書の書式構造に関する情報（書式構造モデル）を用いて入力文書の書式構造の種別を推定する。書式構造情報照合部１６は、推定された書式構造の種別に対応する書式構造モデルと入力文書の構造化特徴の間で、詳細な対応関係を抽出する。未対応・矛盾対応発見修正部１８で対応関係の整合を得た後、文書構造獲得部１９でその対応関係に基づき予め登録されている書式構造モデルに関する情報を入力文書にコピーすることで入力文書の構造及び関連知識を獲得する。



【特許請求の範囲】
【請求項１】　文書より入力画像を生成する画像入力手段と、
入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、
前記画像入力手段により生成された入力画像から幾何学的な図形特徴量を抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された図形特徴量をグループ化して画像特徴を生成し、それぞれの画像特徴間の関係を抽出・管理する特徴構造化手段と、
前記特徴構造化手段で得られた入力画像の画像特徴と、前記書式構造情報登録手段によって予め登録されている処理対象文書の書式構造に関する情報を用いて、入力文書の書式構造の種別の候補を絞りこむ書式構造種別同定手段と、
前記書式構造同定手段で候補となったすべての書式構造モデルと前記特徴構造化手段で構造化された入力文書の特徴との間で対応付けを行ない、最も良く対応づいた書式構造モデルと入力文書の組を選択し、その対応関係を獲得する書式構造情報照合手段と、
前記書式構造情報照合手段で選択された書式構造文書と入力文書における構造化特徴間の対応付けにおいて、不完全な対応付けおよび矛盾した対応付けを解消することにより整合のとれた前記書式構造モデルと入力文書の構造化された特徴間の対応関係を獲得する未対応・矛盾対応発見修正手段と、
前記未対応・矛盾対応発見修正手段によって得られた前記書式構造モデルと入力文書の構造化された特徴間の対応関係に基づいて、予め登録されている当該書式構造モデルに関する情報を入力文書にコピーすることにより入力文書の書式構造と関連情報を獲得する文書構造獲得手段と、
を具備することを特徴とする文書画像処理装置。
【請求項２】　文書より入力画像を生成する画像入力手段と、
前記画像入力手段により生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに前記入力画像における文字成分以外の領域から線分に関する特徴を罫線を構成する図形特徴とみなして抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、
を具備することを特徴とする文書画像処理装置。
【請求項３】　文書より入力画像を生成する画像入力手段と、
前記画像入力手段により生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに前記入力画像における文字成分以外の領域から線分に関する特徴を罫線を構成する図形特徴とみなして抽出する罫線特徴抽出手段と、
前記罫線特徴抽出手段より抽出された罫線に関する図形特徴の集合に対して、交差・接続する罫線を同じグループにまとめることにより表に関する特徴を抽出する表特徴抽出手段と、
を具備することを特徴とする文書画像処理装置。
【請求項４】　文書より入力画像を生成する画像入力手段と、
入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、
前記画像入力手段によって生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴とみなして抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、
前記特徴構造化手段により得られた前記入力文書の表に関する特徴と、予め前記書式構造情報登録手段により登録されている書式構造モデルを構成する表に関する特徴との間で照合処理を行い、表間対応関係を獲得する表照合手段と、
前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線と、同罫線に対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、
前記照合処理結果に基づき特徴間の対応付きの程度を表す照合度を計算し、正しい対応付けが行なわれているか否かの判断を行なう照合結果判定手段と、
を具備することを特徴とする文書画像処理装置。
【請求項５】　文書より入力画像を生成する画像入力手段と、
入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、
前記画像入力手段によって生成された入力画像から幾何学的な図形特徴量を抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された図形特徴量をグループ化して画像特徴を生成し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、
前記特徴構造化手段で得られた入力画像の構造化された画像特徴と、予め前記書式構造情報登録手段によって登録されている処理対象文書の書式構造に関する情報を用

いて、類似度を計算し、最も類似度の高い書式構造モデルあるいは類似度の高いものから順に複数個の書式構造モデルあるいはある一定値以上の類似度を有する書式構造モデルを選び、前記入力文書の書式構造の種別を一つあるいは複数個の候補に絞りこむ書式構造種別同定手段と、
を具備することを特徴とする文書画像処理装置。
【請求項６】　文書より入力画像を生成する画像入力手段と、
入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、
前記画像入力手段により得られた入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴とみなして抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、
前記特徴構造化手段で得られた入力画像の構造化された特徴と、予め前記書式構造情報登録手段によって登録されている処理対象文書の書式構造に関する情報を用いて、類似度を計算し、最も類似度の高い書式構造モデルあるいは類似度の高いものから順に複数個の書式構造モデルあるいはある一定値以上の類似度を有する書式構造モデルを選び、入力文書の書式構造の種別を一つあるいは複数個の候補に絞りこむ書式構造種別同定手段と、
前記書式構造種別同定手段により選択されたそれぞれの書式構造モデルに対して、
前記特徴構造化手段により得られた入力文書の表に関する特徴と、前記書式構造情報登録手段により登録されている当該書式構造モデルを構成する表に関する特徴との間い照合処理を行ない、表間対応関係を獲得する表照合手段と、
前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線と同罫線に対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、
前記罫線照合手段により得られた対応関係に対して、特徴間の対応付きの程度を表す照合度を計算する照合度計算手段と、
前記照合度計算手段で計算されたそれぞれの書式構造モデルの照合度の中から最大照合度を示す書式構造モデルを抽出する照合結果出力手段と、
前記照合結果出力手段によって抽出された書式構造モデルの最大照合度を用いて、入力文書と当該書式構造モデルの構造化特徴間で正しい対応付けが行われているか否かの判定を行なう照合結果判定手段と、
を具備することを特徴とする文書画像処理装置。
【請求項７】　前記書式構造情報登録手段は、
入力文書の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報を登録する際に、正立した処理対象文書の書式構造に関する情報を複数の所定角度で回転させたものを発生させ、それぞれに正立したものから何度回転しているかに関する情報を付与し、それらすべてを処理対象文書の書式構造に関する情報として登録することを特徴とする請求項１または請求項４または請求項５または請求項６記載の文書画像処理装置。
【請求項８】　文書より入力画像を生成する画像入力手段と、
入力文書の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報を登録する際に、正立した処理対象文書の書式構造に関する情報を複数の所定角度で回転させたものを発生させ、それぞれに正立したものから何度回転しているかに関する情報を付与し、それらすべてを処理対象文書の書式構造に関する情報として登録する書式構造情報登録手段と、
前記画像入力手段により生成された入力画像から幾何学的な図形特徴量を抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された図形特徴量をグループ化することで画像特徴を生成し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、
前記特徴構造化手段で得られた入力画像の構造化された画像特徴と、前記書式構造情報登録手段によって予め登録されている処理対象文書の書式構造に関する情報を用いて、入力文書の書式構造の種別を一つあるいは複数個の候補に絞りこむ書式構造種別同定手段と、
前記書式構造種別同定手段で候補となったすべての書式構造モデルと入力文書の前記特徴構造化手段で構造化された特徴との間で対応付けを行ない、最も良く対応づいた書式構造モデルと入力文書の組を選択し、その対応関係を獲得するモデル照合手段と、
前記モデル照合手段で選択された書式構造文書と入力文書における構造化特徴間の対応付けにおいて、不完全な対応付けおよび矛盾した対応付けがなされているか否かを発見し、それらを解消することにより整合のとれた前記書式構造モデルと入力文書の構造化された特徴間の対応関係を獲得する未対応、矛盾対応発見・修正手段と、
前記モデル照合手段で入力画像に対応付いた書式構造モデルが所定角度で回転させたものである場合には、その回転角度を正立する方向に入力画像を回転し、正立した当該書式モデルと対応付ける画像回転手段と、
前記書式構造モデルと入力文書の構造化された特徴間の対応関係に基づいて、予め登録されている当該書式構造モデルに関する情報を入力文書にコピーすることにより入力文書の書式構造と関連情報を獲得する文書構造獲得手段と、
を具備することを特徴とする文書画像処理装置。

【請求項９】　文書より入力画像を生成する画像入力手段と、
入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、
前記画像入力手段によって生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴（罫線特徴）とみなして抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された罫線特徴をグルーープ化することにより表に関する特徴（表特徴）を抽出し、各表特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、
前記特徴構造化手段により得られた入力文書の表特徴と、予め前記書式構造情報登録手段により登録されている書式構造モデルを構成する表特徴との間で照合処理を行い、表間対応関係を獲得する表照合手段と、
前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線特徴とそれに対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、
前記罫線照合手段により獲得された対応関係に対し、特徴間の対応付きの程度を表す照合度を計算し、正しい対応付けが行なわれているか否かの判定を行なう照合結果判定手段と、
前記照合結果判定手段によって正して対応付けが行われていると判定された入力文書と書式構造モデルの罫線特徴間の対応関係において、入力文書の罫線特徴（入力罫線特徴）に対応付いていない書式構造モデルの罫線特徴（未対応モデル罫線特徴）を抽出する手段と、
前記未対応モデル罫線特徴に対応付くべき入力罫線特徴が他のモデル罫線特徴に対応付いている場合には、その対応関係を解消し、未対応モデル罫線特徴とその入力罫線特徴を対応付ける手段と、
前記未対応モデル罫線特徴に対応付くべき欠損した未対応の入力罫線特徴がある場合には、未対応モデル罫線特徴と当該入力罫線特徴を対応付ける手段と、
前記未対応モデル罫線特徴に対応付くべき入力罫線特徴が見つからない場合には、対応付くべき入力罫線特徴を新たに発生させ、未対応モデル罫線特徴と新たに発生させた入力罫線特徴を対応付ける手段によって未対応および矛盾対応を修正する手段と、
を具備することを特徴とする文書画像処理装置。
【請求項１０】　文書より入力画像を生成する画像入力手段と、
入力画像の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報（書式構造モデル）を予め登録する書式構造情報登録手段と、
前記画像入力手段により生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに入力画像における文字成分以外の領域から線分に関する特徴を罫線に関する図形特徴（罫線特徴）とみなして抽出する特徴抽出手段と、
前記特徴抽出手段より抽出された罫線特徴をグループ化することにより表に関する特徴（表特徴）を抽出し、各表特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、
前記特徴構造化手段により得られた入力文書の表特徴と、予め書式構造情報登録手段により登録されている書式構造モデルを構成する表特徴との間で照合処理を行い、表間対応関係を獲得する表照合手段と、
前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線特徴とそれに対応付く書式構造モデルの表を構成する罫線特徴との間の対応関係を獲得する罫線照合手段と、
前記罫線照合手段で獲得された罫線特徴間の対応関係において、当該書式構造モデルの罫線特徴集合における罫線間の接続関係に基づいて、対応する入力文書の罫線特徴集合における罫線間の接続関係を修正する罫線照合後処理手段と、
を具備することを特徴とする文書画像処理装置。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、例えば帳票などの文書上に記載された文字の読みとり、データベースへの自動入力、帳票画像の自動ファイリングに用いられる文書画像処理装置に関するものである。
【０００２】
【従来の技術】近年、書類形態で管理・利用されていた文書を電子化して計算機に入力し、多用な用途で活用している。様々な業務処理で用いられている文書には、表形式のものが多く、文書中の特定の位置に記載されている所望の文字列を所望の形式で計算機に効率よく入力し、管理したいという要求が高まっている。これらの期待に答えるためにこれまでに光学的文字読みとり装置が開発され、実用化されてきた。
【０００３】このような帳票上の特定の位置に記載されている文字列を読み取り対象とする場合、帳票の書式構造を事前に知る必要がある。帳票のように定型的な書式構造を持つ文書に対する文字列の読み取り処理では、書式構造を事前に覚えさせておき、それに基づいて効率良く文字列領域を特定し、読み取るようにしている。
【０００４】このようなアプローチをとるものとしては、対象文書の書式構造に関する知識とそれを利用する処理を分離することで書式構造の多用性への拡張性を高めている。これらのものは書式構造に関する知識として文書構造の物理情報、例えば位置、大きさ、幾何学的関

係などの情報を用いている。代表的な研究として、文献「信学論（Ｄ）、Ｊ71－Ｄ、10、pp. 2050-2058 (1988-10）」と文献「信学論（Ｄ－II）、Ｊ72－Ｄ－II、７、pp. 1029-1039 (1989-07）」がある。

【０００５】一方、最近になって、文書構造に関する物理情報を用いるだけでは対応できない帳票に対しても文字列領域の特定を可能にすることが考えられている。この場合、文字領域を空間的な隣接・接続関係に基づいた論理情報で表現し、識別することを可能にしている。

【０００６】このようなアプローチをとるものとしては、種々の帳票文書の書式構造に関する構成規則を一般化してメタ知識と表現されたものを用いて対象文書画像の書式構造を認識している。代表的な研究として文献「信学論（Ｄ－II）、Ｊ76－Ｄ－II、３、pp. 534-545 (1993-03）」がある。

【０００７】

【発明が解決しようとする課題】文献「信学論（Ｄ）、Ｊ71－Ｄ、10、pp. 2050-2058 (1988-10）」や文献「信学論（Ｄ－II）、Ｊ72－Ｄ－II、７、pp. 1029-1039 (1989-07）」のようなアプローチをとる手法では、物理情報に依存しているので文書が大幅にずれて入力されたり、拡大・縮小によるスケール変換を受けている場合には対応できない。また、文献「信学論（Ｄ－II）、Ｊ76－Ｄ－II、３、pp. 534-545 (1993-03）」のようなアプローチをとる手法では、入力文書は高品質で画像の劣化がないことを前提としているため、処理対象文書の品質が悪く画像情報が不足している場合には所望の処理結果を得られないという問題がある。

【０００８】また、何れの手法においても、以下の（１）～（６）の問題がある。

【０００９】（１）　表が複数混在している場合には対応できない。

【００１０】（２）　表が分裂している場合には対応できない。

【００１１】（３）　罫線がかすれていたり、欠落している場合には対応できない。

【００１２】（４）　罫線分布が局所的に変動している場合には対応できない。

【００１３】（５）　入力された文書の処理方法を限定している（左右に９０度および１８０度回転している文書に対応できない）。

【００１４】（６）　種々の書式構造の文書を一括して読み取らせる場合、適用すべき書式構造モデルの自動同定ができない。

【００１５】本発明は上記事情に鑑みてなされたもので、表形式の帳票などの書式構造を正確に認識でき、効率の良い文字列の領域の特定を可能にした文書画像処理装置を提供することを目的とする。

【００１６】

【課題を解決するための手段】本発明は、文書より入力画像を生成する画像入力手段と、入力文書の書式構造を認識するために用いられる処理対象文書の書式構造に関する情報を登録する際に、正立した処理対象文書の書式構造に関する情報を複数の所定角度で回転させたものを発生させ、それぞれに正立したものから何度回転しているかに関する情報を付与し、それらすべてを処理対象文書の書式構造に関する情報として登録する書式構造情報登録手段と、前記画像入力手段により生成された入力画像から線分と文字成分に関する図形特徴を抽出し、さらに前記入力画像における文字成分以外の領域から線分に関する特徴を罫線を構成する図形特徴とみなして抽出する特徴抽出手段と、前記特徴抽出手段より抽出された罫線に関する図形特徴をグループ化することにより表に関する特徴を抽出し、各表に関する特徴において罫線が交差・接続する部分に生じる接合部に関する情報を抽出し、それぞれの特徴間の関係を抽出・管理する特徴構造化手段と、前記特徴構造化手段で得られた入力画像の構造化された画像特徴と、予め前記書式構造情報登録手段によって登録されている処理対象文書の書式構造に関する情報を用いて、類似度を計算し、最も類似度の高い書式構造モデルあるいは類似度の高いものから順に複数個の書式構造モデルあるいはある一定値以上の類似度を有する書式構造モデルを選び、前記入力文書の書式構造の種別を一つあるいは複数個の候補に絞りこむ書式構造種別同定手段と、前記特徴構造化手段により得られた入力文書の表に関する特徴と、前記書式構造情報登録手段により登録されている当該書式構造モデルを構成する表に関する特徴との間い照合処理を行ない、表間対応関係を獲得する表照合手段と、前記表照合手段により得られた表の対応関係において入力文書の表を構成する罫線と同罫線に対応付く書式構造モデルの表を構成する罫線との間の対応関係を獲得する罫線照合手段と、前記照合処理結果に基づき特徴間の対応付きの程度を表す照合度を計算し、正しい対応付けが行なわれているか否かの判断を行なう照合結果判定手段とにより構成されているモデル照合手段と、前記モデル照合手段で選択された書式構造文書と入力文書における構造化特徴間の対応付けにおいて、不完全な対応付けおよび矛盾した対応付けを解消することにより整合のとれた前記書式構造モデルと入力文書の構造化された特徴間の対応関係を獲得する未対応・矛盾対応発見修正手段と、前記モデル照合手段で入力画像に対応付いた書式構造モデルが所定角度で回転させたものである場合には、その回転角度を正立する方向に入力画像を回転し、正立した当該書式モデルと対応付ける画像回転手段と、前記書式構造モデルと入力文書の構造化された特徴間の対応関係に基づいて、予め登録されている当該書式構造モデルに関する情報を入力文書にコピーすることにより入力文書の書式構造と関連情報を獲得する文書構造獲得手段とを具備し、この文書構造獲得手段により得られる結果に基づいて入力画像の書式構造を

認識するように構成されている。

【００１７】

【作用】この結果、本発明は、入力文書画像から罫線に関する図形特徴量と文字成分に関する図形特徴量を抽出し、これらのかかわり合いから文字成分以外の領域で正確に罫線の特徴量を抽出できる。また、この罫線特徴をグループ化することにより表に関する特徴量が得られ、さらに各表において罫線に関する特徴とそれらの交差・接合により生じる接合部を抽出し、全体ー部分関係で記述・管理することにより図形特徴量に対してより豊富な情報を付加することができる。それらを効率的に検索することができる。これにより文書内に表が複数個混在していても各々の情報を抽出することができ、また罫線に関する特徴も表ごとに区別することができる。

【００１８】本発明では、入力画像から得られた構造化された特徴と予め登録されている書式構造モデルの構造化特徴との間で照合処理を行なうことにより入力文書を解釈するが、照合処理の前に書式構造モデルの種別を同定することで登録されているすべての書式構造モデルとの照合処理を行なうことを避けることができる。この処理では、文書中に含まれている接合部の数や罫線の数を用いているので入力文書と書式構造モデルの間で拡大縮小に伴うスケール変換がなされていたり、表や罫線の構成要素の大きさが部分的に変動していても安定した結果を得ることができる。書式構造モデルの種別を同定する時に唯一の結果を出力するのではなく複数個の候補を出力することで同定誤りが生じないようにしている。候補となっている書式構造モデルの構造化特徴と入力文書の構造化特徴との間で行なわれる照合処理では、もっとも良く対応付く書式構造モデルを選ぶことができ、得られた対応関係が正しく獲得されているか否かの判断を行なうことにより処理結果の信頼性を高めている。このとき、照合処理まず表単位に行なわれ、次いで罫線単位に行なわれる。この結果、局所的にまったく同じ特徴量を持つ罫線でもそれが所属する表が異なれば誤った対応付けが行なわれることはない。

【００１９】また、本発明では、対応関係に対してどちらかの構造化特徴の欠落による不完全な対応付きや矛盾した対応付きの有無およびこの箇所を発見し、それを解消する。この結果、整合のとれた対応関係を獲得することができ、安定した処理結果を得ることができる。この対応関係に基づいて予め登録されている書式構造モデルに関連する情報を入力文書に複写することにより入力文書の文書構造を獲得し、読みとるべき文字列の記載位置を正確に特定することができる。

【００２０】

【実施例】以下、図面を参照して本発明の一実施例を説明する。図１は本実施例に係わる画像処理装置の概略構成を示すブロック図である。

【００２１】本発明は、画像入力部１１、特徴抽出部１２、特徴構造化部１３、書式構造情報登録部１４、書式構造種別同定部１５、書式構造情報照合部１６、未対応・矛盾対応発見修正部１７、照合結果判定部１８、及び文書構造獲得部１９の各機能部から構成されている。

【００２２】また、処理対象とする帳票は、図２に示すように、罫線により文字列領域が規定されているものとする。

【００２３】画像入力部１１は、スキャナ装置やＴＶカメラ、ＦＡＸデータの入力部などによって構成されるものである。画像入力部１１は、処理対象である帳票等の文書の画像を検出してシステム内に取り込む。画像入力部１１によって入力された画像（入力画像）は、特徴抽出部１２に送られる。

【００２４】特徴抽出部１２は、画像入力部１１から得られた入力画像から、罫線や文字成分に関する幾何学的な図形特徴量をそれぞれ基本特徴として抽出する。特徴抽出部１２により抽出された基本特徴の集合は特徴構造化部１３に送られる。

【００２５】特徴構造化部１３は、特徴抽出部１２により抽出された基本特徴の集合について、例えば罫線に関する特徴をグループ化することにより得られる表や、罫線が交差・接続する部分に生じる接合部などに関する情報を構造化特徴として抽出し、さらにそれぞれの特徴間の関係を記述、管理する。

【００２６】書式構造情報登録部１４は、システムとオペレータとの対話的な入力作業により、処理対象文書に関する知識の登録を行なう。本実施例では、処理対象文書の種類ごとに用意された種々の知識の総体をモデルと呼び、モデルを定義したときに用いた文書をモデル文書と呼ぶことにする。モデルは例えば、そのモデル文書の構造化特徴に関する情報などを有している。

【００２７】このとき、書式構造情報登録部１４は、正立している構造化特徴から、右に９０度回転したもの、左に９０度回転したもの、１８０度回転したものの３種類の構造化特徴を発生させ、それぞれが正立したものから何度回転されているものであるかという情報を付与して知識ベースとして格納してもよい。

【００２８】この他に、例えば罫線により規定されている領域内に含まれる文字列を文字認識装置で認識・コード化させるような場合には、各種情報を構造化特徴に関連付けて知識ベースとして登録・管理するようにしてもよい。

【００２９】各種情報には、例えば、文字認識対象領域の指定に関する情報、文字認識対象領域の文字列方向の情報、文字読み取り処理時に拘束条件として働く文字種情報及び筆記形態情報、文字認識後処理のための項目属性情報、文字認識結果の対応関係などの情報がある。

【００３０】書式構造種別同定部１５は、特徴構造化部１３によって抽出された入力画像に対する構造化特徴と、予め書式構造情報登録部１４によって登録されてい

るモデル文書のそれぞれの構造化特徴との間で、類似度計算を行なうことにより入力文書のフォーマットの種別を同定する。すなわち類似度値が高いモデルのフォーマットが入力文書のフォーマットである可能性が高いものと判別する。書式構造種別同定部１５は、該当する構造化特徴を次段の書式構造情報照合部１６に出力する。

【００３１】なお、書式構造種別同定部１５は、最も類似度値の高いモデルのみに注目するのではなく、例えばある事前に定められたしきい値以上の類似度値を示す全てのモデル文書の構造化特徴を候補として、書式構造情報照合部１６に入力構造化特徴と共に送り込むようにしてもよい。

【００３２】書式構造情報照合部１６は、書式構造種別同定部１５によって候補とされた全てのモデル文書と入力文書の間で構造化特徴間の対応関係を獲得する。また、書式構造情報照合部１６は、入力文書の構造化特徴と各モデル文書の構造化特徴との対応付け結果に対して対応付きの度合いを表す尺度（以後、照合度と呼ぶ）を計算する。書式構造情報照合部１６は、構造化特徴間の対応関係を示す情報、及び照合度を示す情報を照合結果判定部１７に出力する。

【００３３】なお、書式構造種別同定部１５から複数のモデル文書の構造化特徴が送られてきた場合には、書式構造情報照合部１６は、さらに最大照合度を示すモデル文書を選択・出力するようになっていてもよい。

【００３４】照合結果判定部１７は、書式構造情報照合部１６によって得られた情報に基づいて、最大照合度を示す文書モデルと入力文書との間で対応関係が取れているか否かを判定する。ここで対応関係が取れていると判定された場合には、後段の未対応・矛盾対応発見修正部１８に着目入力文書とモデル文書間の対応関係に関する情報が送られるようにする。対応関係が取れていないとみなされる場合には、入力文書を棄却して、次の文書を入力するようにオペレータに促す。

【００３５】未対応・矛盾対応発見修正部１８は、書式構造情報照合部１６で獲得された入力文書の構造化特徴とモデル文書の構造化特徴の対応関係において、誤った特徴の抽出や必要な特徴の欠落のために不完全な対応や矛盾した対応が生じているか否かを発見する。未対応・矛盾対応発見修正部１８は、不完全な対応や矛盾した対応を発見した場合には、それらを解消することにより整合のとれた入力・モデル間の対応関係を獲得した上で、文書構造獲得部１９に出力する。

【００３６】文書構造獲得部１９は、未対応・矛盾対応発見修正部１８で得られた整合のとれた入力文書とモデル文書の構造化特徴間の対応関係に基づいて、予め書式構造情報登録部１４で登録されているモデル文書に関する知識を入力文書にコピーすることにより入力文書の文書構造及び関連知識を獲得する。

【００３７】次に、上述した必須構成要素を含む具体的な文書画像処理システムについて説明する。ここでは、図３に示すように、文字認識装置と組み合わせた文書画像処理システムについて説明する。このシステムは、入力文書から得られた入力画像において、特定の位置に記載されている文字列領域を自動的に抽出し、さらにその文字列画像を認識・コード化して、所望の出力様式で出力するという動作をするものである。以後、図２に示す文書（以下、特に断らない限り図２の文書を入力文書と呼ぶ）を用いて説明する。

【００３８】図３に示すように、文書画像処理システムは、画像入力部２１、２値化処理部２３、前処理部２５、特徴抽出部２７、特徴構造化部２９、モデル登録部３１、フォント種別同定部３３、モデル照合部３５、照合結果判定部３７、未対応矛盾対応発見修正部３９、文書構造獲得部４１、文字列領域抽出部４３、文字認識部４５、及び文字認識結果出力部４７によって構成されている。

【００３９】画像入力部２１は、図１において説明した画像入力部１１と同じものとして説明を省略する。

【００４０】２値化処理部２３は、画像入力部２３から取り込まれた文書画像を、公知である２値化処理により白と黒の２値の画像データに変換し、前処理部２５に出力する。

【００４１】前処理部２５は、２値化処理部２３から出力された２値画像について、例えば文献「信学技報、ＰＲＵ９２－３２、１９９２」に記載されている傾き検出・補正処理により、傾きのない２値画像に変換する。さらに、前処理部２５は、傾きが補正された２値画像に対して、公知である黒連結成分抽出処理により、連結する黒画素のまとまりを囲む外接矩形枠を生成し、その大きさ（縦幅と横幅の長さ）や位置座標値を抽出、管理し、その結果を特徴抽出部２７に出力する。

【００４２】なお、入力文書の位置座標は左上端を原点とし、ｘ座標値は右方向に次第に大きくなり、ｙ座標値は下方向に次第に大きくなるように定義されているものとする。本実施例で触れる文書画像は全てこの座標系で定義されている。この内、縦幅と横幅の長さが、それぞれ動的に検出されたしきい値th1 、th2 よりも小さく、かつ最近傍の他の黒連結成分矩形から距離th3 以上離れているものを、ノイズあるいは網点であるとして除去するようにしてもよい。

【００４３】以後、画像入力部２１から入力された文書画像に対して、２値化処理部２３における２値化処理、前処理部２５における前処理を施して得た画像を入力文書画像または単に入力画像と呼ぶ。入力画像は、特徴抽出部２７に送られる特徴抽出部２７は、前処理部２５から得られた入力画像から幾何学的な図形特徴を抽出するものであり、図４に示すように、線分抽出部２７ａ、文字候補矩形抽出部２７ｂ、及び罫線特徴抽出部２７ｃによって構成されている。特徴抽出部２７は、以下に示す

手順で幾何学的な図形特徴を抽出する。

【００４４】まず、線分抽出部２７ａは、例えば、以下に述べる手順に基づいて入力画像から線分を抽出する。このとき線分抽出処理は、垂直方向と水平方向の２つの方向に限定して実施されるようにしてもよい。具体的な例として、垂直方向の線分を抽出する場合について説明する。なお、水平方向の線分についても同様の手順により抽出することができる。

【００４５】Ｓｔｅｐ１：文書画像を垂直方向に順次走査し、各走査線において、予め定めた例えば３ドット以上の長さを持つ連続する黒画素の連なりの集合（ＢＬ＝｛ｂｌｉ｜ｉ＝１，２，…，ｎ｝）を抽出する。

【００４６】Ｓｔｅｐ２：ＢＬの要素のうち、水平方向に隣接しているものを統合してまとめることにより、さらに、その集合（ＢＬＧ＝｛ｂｌｇｊ｜ｊ＝１，２，…Ｍ｝）を抽出する。

【００４７】Ｓｔｅｐ３：ＢＬＧの各要素に含まれるＢＬにおいて、垂直方向に最も長い黒画素の連なりｂｌｍａｘ を抽出し、その長さｂｌmax のα（例えばα＝０．３）倍未満の長さを持つｂｌｉ を削除する。

【００４８】Ｓｔｅｐ４：残ったｂｌｉ を内接する矩形を抽出し、その集合（ＲＬ＝｛ｒｌｋ｜ｋ＝１，２，…，Ｐ｝）を抽出する（このとき、ｒｌｋ とそれを構成するｂｌｉ とを相互に関連づける）。ｒｌｋ の左上端および右下端のｙ座標値を検出し、それぞれｒｌｋ ｙ１、ｒｌｋ ｙ２ とする。

【００４９】Ｓｔｅｐ５：ＲＬの各要素を水平方向（左から右への方向）に順次走査し、各走査線において最初に出現するｂｌｉ のｘ座標（ｘis）と最後に出現するｂｌｉのｘ座標（ｘie）を保持する。

【００５０】Ｓｔｅｐ６：各ｒｌｋ の各水平走査線において得られたｘisの集まりと、ｅieの集まりの平均値ｒｌｋ ｘ１，ｒｌｋ ｘ２ をそれぞれ計算する。

【００５１】Ｓｔｅｐ７：各ｒｌｋ の左上端および右下端の座標をそれぞれ（ｒｌｋ ｘ１，ｒｌｋ ｙ１）、（ｒｌｋ ｘ２，ｒｌｋ ｙ２）に設定し、それを線分として抽出する。

【００５２】線分抽出部２７ａにより、入力画像から抽出された線分の例を図５に示している。図５に示すように、この段階では文字の部分において、文字を構成する短い線分を検出している。これらを取り除くように以下の処理を行なってもよい。すなわち、線分抽出部２７ａによる線分抽出処理に前後して、あるいは同時に黒連結成分群に対して、文字候補矩形抽出部２７ｂは、以下に述べる手順により、文字とみなすことのできる黒連結成分（以後、文字候補矩形と呼ぶ）を選出する。

【００５３】Ｓｔｅｐ１：各黒連結成分を内接する矩形を抽出し、その縦幅ｃｈと横幅ｃｗを求める。

【００５４】Ｓｔｅｐ２：ｃｈとｃｗの各値に対して出現頻度を求め、最頻値を抽出する。最頻値を示すｃｈを入力文書中の文字の平均的な文字高さ（ＣＨ）、最頻値を示すｃｗを入力文書中の文字の平均的な文字幅（ＣＷ）とみなす。

【００５５】Ｓｔｅｐ３：（ＣＷ－ｔｈ）≦ｃｗ≦（ＣＷ＋ｔｈ）かつ（ＣＨ－ｔｈ）≦ｃｈ≦（ＣＨ＋ｔｈ）を満たす黒連結成分を文字候補矩形として抽出する。

【００５６】文字候補矩形抽出部２７ｂにより抽出された文字候補矩形の抽出結果例を図６（文字部を外枠矩形で囲んだ例）に示している。

【００５７】なお、文字候補矩形抽出処理を線分抽出処理より先に行ない、得られた文字候補矩形以外の入力画像上の部分に対して垂直方向と水平方向の両方向に、以下に述べるフィルタリング処理を適用して、罫線がかすれていたり、とぎれているような場合でも線分抽出処理結果を安定させるようにしてもよい。

【００５８】この場合、特徴抽出部２７は、図７に示すように構成される。すなわち、文字候補矩形抽出部２７ｂの処理結果がフィルタリング処理部２７ｄに出力されフィルタリング処理が施される。そして、フィルタリング処理の後に、線分抽出部２７ａによる線分抽出処理実行される。

【００５９】フィルタリング処理部２７ｄによるフィルタリング処理は、例えば、２値画像において水平（垂直）方向の走査線上のある決まった長さ以内で連続する白画素を全て黒画素に置き換えるという処理で実現される。これにより、かすれやとぎれのある罫線部分が補正されるので、線分抽出処理結果が安定される。

【００６０】罫線特徴抽出部２７ｃは、線分抽出部２７ａによって抽出された線分群において、図８に示すような文字候補矩形に交差あるいは包含されている線分を除去し、残った線分を罫線特徴として抽出する。

【００６１】この時点では、各罫線特徴は、入力画像における当該画像を囲む外接矩形の左上端と右下端の位置座標、外接矩形の縦幅及び横幅などの種々の情報で表現されるようにしてもよい。また、罫線特徴に「対応付いた罫線特徴の識別番号」を用意しておいて、後段のモデル照合部３５で実施される照合処理において、対応付いた相手の識別番号を格納するようにしてもよい。

【００６２】得られた罫線特徴の集合は、さらに水平罫線と垂直罫線の２種類に分類され、水平罫線特徴の集合とその要素数、垂直罫線特徴の集合とその要素数が抽出される。

【００６３】入力画像に対する罫線特徴抽出結果の例を図９に示す。ここで、入力画像から抽出された水平罫線の集合（水平罫線特徴集合）と垂直罫線の集合（垂直罫線特徴集合）の各要素に、

ＩＨＬ＝（ｉｈｌ１，ｉｈｌ２，ｉｈｌ３，ｉｈｌ４，ｉｈｌ５，ｉｈｌ６，ｉｈｌ７），

ＩＶＬ＝（ｉｖｌ１，ｉｖｌ２，ｉｖｌ３，ｉｖｌ４，ｉｖｌ５，ｉｖｌ６，ｉｖｌ７，ｉｖｌ８，ｉｖｌ９）

となるように識別番号を付与する。

【００６４】また、水平罫線特徴集合の要素数をｉｈ-num、垂直罫線特徴集合の要素数をｉｖ-numとし、水平罫線特徴集合と垂直罫線特徴集合をまとめて入力罫線特徴集合と呼ぶことにする。入力罫線特徴集合は、後段の特徴構造化部２９に送られる。以後の説明において上記記号を用いることにする。

【００６５】特徴構造化部２９は、特徴抽出部２７から得られた入力罫線特徴集合から以下に述べる手順で罫線に関する構造化特徴を抽出するもので、図１０に示すように、罫線グループ化処理部２９ａ、表特徴抽出部２９ｂ、罫線接合部検出部２９ｃ、及び特徴間関係記述部２９ｄによって構成されている。

【００６６】まず、罫線グループ化処理部２９ａは、交差・接続する罫線特徴を一まとめにグループ化する。罫線グループ化処理アルゴリズムは例えば以下のようになる。このアルゴリズムは、交差・接続する罫線には同じラベルを与えることにより罫線をグループ化するものである。

【００６７】Ｓｔｅｐ１：ラベル番号を初期化する。

【００６８】Ｓｔｅｐ２：１本の垂直罫線を選択する。

【００６９】Ｓｔｅｐ３：ステップ２で選択された当該垂直罫線に直交する全ての水平罫線を抽出する。

【００７０】Ｓｔｅｐ４：当該垂直罫線か当該水平罫線のいずれかに既にラベルが付与されている場合には、その中で最小値のラベルを当該罫線の全てに付与する。

【００７１】Ｓｔｅｐ５：当該垂直罫線と当該水平罫線のいずれにもラベルが付与されていない場合には、それら全てに新しいラベル番号を付与し、ラベル番号を更新する。

【００７２】Ｓｔｅｐ６：Ｓｔｅｐ２からＳｔｅｐ５を全ての垂直罫線に適用する。（Ｓｔｅｐ２からＳｔｅｐ６までを手続きＡとする）

Ｓｔｅｐ７：１本の水平罫線を選択する。

【００７３】Ｓｔｅｐ８：当該水平罫線に直交する全ての垂直罫線を抽出する。

【００７４】Ｓｔｅｐ９：当該水平罫線か当該垂直罫線のいずれかに既にラベルが付与されている場合にはその中で最小値のラベルを当該罫線の全てに付与する。

【００７５】Ｓｔｅｐ10：当該垂直罫線と当該水平罫線のいずれにもラベルが付与されていない場合には、それら全てに新しいラベル番号を付与し、ラベル番号を更新する。

【００７６】Ｓｔｅｐ11：Ｓｔｅｐ7 からＳｔｅｐ10を全ての垂直罫線に対して適用する（Ｓｔｅｐ7 からＳｔｅｐ11までを手続きＢとする）。

【００７７】Ｓｔｅｐ12：手続きＡと手続きＢをラベル番号の更新がなくなるまで繰り返す。

【００７８】Ｓｔｅｐ13：同じラベルが付与されている罫線特徴をグループ化する。

【００７９】例えば、図９に示す入力罫線特徴集合に対して、前述のようなグループ化処理を適用すると以下に示す２つのグループが得られる。

【００８０】Group１：（ｉｈｌ1，ｉｈｌ2，ｉｈｌ3，ｉｈｌ4，ｉｈｌ5，ｉｖｌ1，ｉｖｌ2，ｉｖｌ3，ｉｖｌ4，ｉｖｌ5，ｉｖｌ6），
Group２：（ｉｈｌ6，ｉｈｌ7，ｉｖｌ7，ｉｖｌ8，ｉｖｌ9）

表特徴抽出部２９ｂは、罫線グループ化処理部２９ａによって得られたグループごとに、グループに含まれる罫線を内接する矩形（以後、表枠と呼ぶ）を抽出する。さらに、表特徴抽出部２９ｂは、表特徴として、例えば左上端の位置座標、右下端の位置座標、表の縦幅、表の横幅、重心の位置座標、水平罫線数、当該表に含まれる水平罫線の集合、垂直罫線数、当該表に含まれる垂直罫線の集合、当該表に含まれる接合部数、他の表に内接されているか否かの情報などを抽出する。この時、水平（垂直）罫線特徴は、その左上端のｙ（ｘ）座標値の昇順にソートされていてもよい。

【００８１】ここで接合部数については、後段の罫線接合部検出部２９ｃで抽出され、表特徴として格納される。例えば、Group１からは、図１１に示すように以下の表特徴ｉｔ1 が得られる。また、Group２から得られる表特徴をｉｔ2 とする。

【００８２】

表特徴：ｉｔ1

左上端の位置座標	（ｉｔ1 ｘ1，ｉｔ1 ｙ1）
右下端の位置座標	（ｉｔ1 ｘ2，ｉｔ1 ｙ2）
表の縦幅	ｉｔ1 height
表の横幅	ｉｔ1 width
重心の位置座標	（ｉｔ1 ｃｘ，ｉｔ1 ｃｙ）
水平罫線数	ｉｈｌ1 num
当該表に含まれる水平罫線の集合	ＩＨＬ1
垂直罫線数	ｉｖｌ1 num
当該表に含まれる垂直罫線の集合	ＩＶＬ1
当該表に含まれる接合部数	ｉ1 junc-num
他の表に内接されているか否かの情報	nest flag

it1 height＝it1 y2 －it1 y1 ＋1，
it1 width ＝it1 x2 －it1 x1 ＋1，
it1 cx＝（it1 x1 ＋it1 x2 ）／2，
it1 cy＝（it1 y1 ＋it1 y2 ）／2，
ihl1 num ＝5，
IHL1 ＝（ihl1，ihl2，ihl3，ihl4，ihl5），
ivl1 num ＝6，
IVL1 ＝（ivl1，ivl2，ivl3，ivl4，ivl5，ivl6），
nest flag ＝0（すなわち他の表に含まれていない）

１枚の入力帳票に複数の表が含まれていることを考慮して、さらにページ特徴を抽出、管理する。ページ特徴は、例えば、表数、表特徴の集合、水平罫線数、垂直罫線数、及び接合部数によって定義される。

【００８３】例えば、図９の入力罫線集合からは、ページ特徴：ＩＰとして、
表数＝２、
表特徴の集合ＩＴ＝（it1，it2）、
水平罫線数ihl-num＝７、垂直罫線数ivl-num＝９、接合部数ijunc-num＝２８、が得られる。

【００８４】次に、罫線接合部検出部２９ｃは、各表において、そこに含まれる水平罫線と垂直罫線の交差部分・接続部分（以後、接合部と呼ぶ）を求め、さらに各表特徴で接合部の個数を管理する。罫線接合部検出部２９ｃの動作手順は例えば次のようになる。

【００８５】Ｓｔｅｐ１：一本の水平罫線を選択する。

【００８６】Ｓｔｅｐ２：着目した水平罫線に交差・接続する全ての垂直罫線を抽出する。

【００８７】Ｓｔｅｐ３：当該水平罫線と当該垂直罫線の交点を求め、水平罫線ごとにその座標値を管理する。各水平罫線では接合部特徴はｘ座標値の昇順にソートしておく。

【００８８】Ｓｔｅｐ４：Ｓｔｅｐ1 からＳｔｅｐ3 までを全ての水平罫線に対して実施する。

【００８９】Ｓｔｅｐ５：一本の垂直罫線を選択する。

【００９０】Ｓｔｅｐ６：着目した垂直罫線に交差・接続する全ての水平罫線を抽出する。

【００９１】Ｓｔｅｐ７：当該垂直罫線と当該水平罫線の交点を求め、垂直罫線ごとにその座標値を管理する。各垂直罫線では接合部特徴はｙ座標値の昇順にソートしておく。

【００９２】Ｓｔｅｐ８：Ｓｔｅｐ５からＳｔｅｐ７までを全ての垂直罫線に対して実施する。

【００９３】例えば、表特徴ｉｔ1 では、水平罫線集合ＩＨＬ1 と、垂直罫線集合ＩＶＬ2から、図１２に示す接合部が得られ、その接合部数「２２」を表特徴に加える（i1 junc-num＝２２とする）。

【００９４】特徴間関係記述部２９ｄは、以上の処理結果より得られた情報を、例えば図１３のように関係づけて管理する。この結果、ページ特徴から表特徴、罫線特徴、接合特徴と階層的に関連づけられて管理され、特徴に関する情報を効率的に検索できるようになる。

【００９５】以後、これらの特徴を総称して構造化特徴と呼ぶ。また、入力文書から抽出された構造化特徴を入力構造化特徴、モデル文書から抽出された構造化特徴をモデル構造化特徴と呼ぶ。

【００９６】一方、処理対象文書に対する処理とは別に、モデル登録部３１は、オペレータによって提示された処理対象文書の種類ごとの一例であるサンプル文書をもとに、オペレータとの間の対話的入力作業によりモデルの登録を行なう。ここでは、モデル登録作業時のモデル登録部３１の動作の一例について説明する。

【００９７】まず、画像入力部２１を介して登録対象のサンプル文書を入力し、２値化処理部２３、前処理部２５を経て文書画像（以後、モデル画像と呼ぶ）を取得する。次いで、特徴抽出部２７においてモデル画像から罫線特徴を抽出する。

【００９８】モデル登録部３１は、特徴抽出部２７における抽出結果に対して特徴抽出処理の誤りを推定し、それを修正する旨のメッセージを、表示部３１ａのディスプレイの画面上に表示して、修正作業をオペレータに指示する。

【００９９】特徴抽出処理の誤りの推定方式は、例えば、端点が接合部となっていない罫線を見つけて、その罫線が正しく抽出されていないとみなすことにより実現できる。モデル登録部３１は、表示部３１ａに表示された指示に従いオペレータによって画面上で修正された罫線特徴を入力部３１ｂを介して入力する。

【０１００】修正作業が完了すると、修正された罫線特徴は特徴構造化部２９に送られ、構造化特徴が抽出される。この構造化特徴は、未知入力文書を処理する際に、フォーマット種別同定部３３、モデル照合部３５、未対応・矛盾対応発見修正部３９、照合結果判定部３７などで用いられる。

【０１０１】抽出・修正された罫線特徴は、画面上にモデル画像と重ねて表示される。オペレータは、書式構造情報登録部３１が画面上に出力するメッセージに従いながら、モデル文書に対応付くべき入力文書画像を処理するために必要な知識を構築していく。

【０１０２】この結果、モデルとして構造化特徴の情報、文字認識対象領域の指定に関する情報、文字認識対象領域の文字列方向の情報、文字読み取り処理時に拘束条件として働く文字種情報、筆記形態情報、文字認識後処理のための項目属性情報、文字認識結果の対応関係などの情報が知識ベース３１ｃに格納される。

【０１０３】このうち構造化特徴は、正立した画像に対して作られている。そこで、モデル登録部３１は、知識ベース３１ｃへの登録時に、左に９０度回転したものと右に９０度回転したものと１８０度回転したものをそれ

ぞれ発生させ、さらに何度回転されているかを示す情報を付加するようにしてもよい。

【０１０４】これら４方向の構造化特徴と入力構造化特徴は、後段のモデル照合部３５において、それぞれの方向の特徴と入力画像とを照合させることにより、入力文書の文書方向が未知であっても、モデルマッチングで最良マッチングした構造化特徴に付与されている回転角度を知ることができる。従って、その角度を０度にする方向に入力画像を回転させれば、必ず正立した入力画像を得ることができる。

【０１０５】モデル登録部３１は、以上の処理を全てのモデル文書に対して行なう。ここでは、一例として図１４と図１５に示す２種類の構造化特徴がモデルとして登録されたものとする（以後、図１４をモデル１、図１５をモデル２と呼ぶ）。

【０１０６】フォーマット種別同定部３３は、モデル登録部３１によって予め登録されているモデルから、入力文書に対応するもの（あるいは対応する可能性のあるもの）を選出する。

【０１０７】ここでは、入力文書の構造化特徴と登録されている全てのモデルの構造化特徴との間で類似度計算を行ない、最も類似度の高いモデルあるいは、ある一定の値以上の類似度を有するモデルの構造化特徴を後段のモデル照合部３５に送り込む。

【０１０８】類似度計算に用いる特徴としては、例えば文書中に含まれる、特徴構造化２９によって抽出された接合部数を用いる。接合部数という特徴は、未知入力文書が寸法の拡大・縮小、さらには各表が独立してスケール変換されている場合にも、影響を受けない特徴であるので書式構造種別同定のための類似度計算に適している。

【０１０９】この他に、文書中に含まれる水平罫線特徴数や垂直罫線特徴数、表特徴数なども類似度計算のための特徴として有効である。ここでモデル文書は適宜に順序づけがなされているとする。

【０１１０】また、類似度は以下のようにして求まる。ここで、モデルの総数：model-num、入力文書中の接合部数：ｉｊｎ、ｋ番目のモデル文書中の接合部数：ｍｊｎｋ（ただし１≦ｋ≦model-num）、入力文書とｋ番目のモデル文書との類似度：ｓｉｍｋ（ただし－100≦ｓｉｍｋ≦100）とすると、類似度は例えば、

ｉｊｎ＞０またはｍｊｎｋ＞０のとき

$sim_k = 100 - 200 \times |ijn - mjn_k| / (ijn + mjn_k)$

ｉｊｎ＝０かつｍｊｎｋ＝０のとき

$sim_k = -100$

により求まる。

【０１１１】得られた類似度のうち類似度が最大であるもの、あるいは予め設定されたしきい値th5 以上のものを候補のモデルとして入力文書の構造化特徴とともにモデル照合部３５に送る。このとき、全てのモデルにおいて類似度がしきい値th5 未満である場合、以後の処理を中断して、入力文書を棄却してもよいし、また、全てのモデルの構造化特徴をモデル照合部に送り込むようにしてもよい。

【０１１２】例えば、登録されているモデル文書が図１４と図１５の２種類である場合、入力文書とそれら２つのモデル文書との類似度計算は以下のようになる。ここで入力文書の接合部数：ｉｊｎ＝28、モデル１の接合部数：mjn_1 ＝31、モデル２の接合部数：mjn_2 ＝26とし、入力文書とモデル１との類似度をsim_1、入力文書とモデル２との類似度をsim_2、しきい値：th5＝80とする。

【０１１３】

$sim_1 = 100 - 200 \times |28-31| / (28+31) = 89.83$

$sim_2 = 100 - 200 \times |28-26| / (28+26) = 92.59$

以上の計算結果からsim_1とsim_2が共にしきい値th5 を越えているので、双方とも入力文書に対応付く可能性のあるモデルとしてそれらの構造化特徴を後段のモデル照合部３５に送る。

【０１１４】モデル照合部３５は、入力文書から抽出された構造化特徴（以後、入力構造化特徴と呼ぶ）と、書式構造種別同定部で選出された一つあるいは複数個のモデル文書の構造化特徴（以後、モデル構造化特徴と呼ぶ）を受け取り、それらの間で以下に述べる照合処理を行ない、照合度を計算する。

【０１１５】照合処理は、当該入力構造化特徴と当該モデル構造化特徴との間の対応付け処理のことであり、照合度はその対応付きの程度を表す尺度である。ここでは、最も高い照合度を示すモデル文書が入力文書に対応するものであるとし、入力構造化特徴とそのモデル構造化特徴との間の対応関係に関する情報を、後段の照合結果判定部に送り込む。最大照合度を示すモデル文書と入力文書との間の対応関係が獲得されているか否かの最終的な判断は照合結果判定部で行なわれる。

【０１１６】特徴構造化部２９で得られた構造化特徴は、図１３に示すように、ページ特徴から表特徴へ、さらに表特徴から罫線特徴へと全体－部分という階層的な関係が抽出されて構造化されている。この特質を用いる場合、照合処理は階層的に実施される。これに対応して、モデル照合部３５は、さらに図１６に示すように、選択部３５ａ、表照合部３５ｂ、罫線照合部３５ｃ、照合度計算部３５ｄ、及び照合結果出力部３５ｅによって構成される。

【０１１７】まず、選択部３５ａは、複数個のモデルの構造化特徴から一つのモデルの構造化特徴を選択し、入力文書の構造化特徴とともに表照合部３５ｂに送る。

【０１１８】次いで、表照合部３５ｂは、入力文書の表

特徴の集合（以後、入力表特徴集合と呼ぶ）とモデル文書の表特徴の集合（以後、モデル表特徴集合と呼ぶ）との間で対応付けを行なう。

【０１１９】表間対応がとれた場合には、さらに罫線照合部３５ｃは、入力文書の当該表の内部に含まれる罫線特徴（以後、入力罫線特徴と呼ぶ）とモデル文書の当該表の内部に含まれる罫線特徴（以後、モデル罫線特徴と呼ぶ）との間で対応付けを行なう。

【０１２０】罫線間の対応がとれた場合には、照合度計算部３５ｄは、（入力文書と当該モデル文書との間の）照合度を計算する。ここで、表間対応と罫線間対応のどちらも得られなかった場合には、照合度に最低値（－１００）を代入する。

【０１２１】照合結果出力部３５ｅは、それまでに計算されている最も高い照合度を示すモデル文書と入力文書の組を選び、その構造化特徴間の対応関係に関する情報を後段の照合結果判定部３７に出力する。

【０１２２】以下では、フォーマット種別同定部３３から図１４と図１５の２種類のモデルと入力文書の構造化特徴がそれぞれ送られてきた場合のモデル照合部３５の処理動作例について説明する。

【０１２３】２種類のモデルのうち、選択部３５ａによってモデル１が選ばれ、入力文書とともにそれぞれの構造化特徴が表照合部３５ｂに送り込まれたとする。ここで、入力文書の表特徴数をｉｔ-num＝２とし、表特徴集合をＩＴとすると、ＩＴ＝（it_1，it_2）（図９参照）モデル１の表特徴数をmt_1-num ＝２とし、表特徴集合をＭＴ１とすると、ＭＴ１＝（mt_{11}，mt_{12}）（図１４参照）とする。

【０１２４】表照合部３５ｂは、さらに図１７に示すように、対応可能ペア検出部３５ｂ－１、異種対応可能ペア間両立関係判定部３５ｂ－２、及び最良マッチング抽出部３５ｂ－３によって構成されており、次のように動作する。

【０１２５】まず、対応可能ペア検出部３５ｂ－１は、選択部３５ａで選択されたモデルの表特徴集合の各要素に対して、それと対応付く可能性のある入力の表特徴をすべて検出し、対応可能ペアとして管理する。すなわち、

Ｓｔｅｐ１：任意のＭＴから任意の一つの表特徴mt_k（着目モデルの表特徴集合のうちｋ番目の表特徴）を選ぶ。

【０１２６】Ｓｔｅｐ２：表特徴mt_kの接合部数とＩＴの各表特徴の接合部数との間で類似度を計算をする。

【０１２７】Ｓｔｅｐ３：例えばmt_kとit_j（入力表特徴集合のｊ番目の表特徴）の類似度sim_{kj}が、あらかじめ設定したしきい値th6 以上である場合には、対応可能であるとして、その組（mt_k，it_j）を対応可能ペアとして保持する。

【０１２８】Ｓｔｅｐ４：Ｓｔｅｐ３を表特徴集合ＩＴのすべての要素に対して適用する。

【０１２９】Ｓｔｅｐ５：Ｓｔｅｐ１～Ｓｔｅｐ４までを着目ＭＴのすべての要素に対して適用する。

【０１３０】ここで、it_jの接合部数を$ijnj$、mt_kの接合部数をmjn_kとすると、類似度は、例えば次の式で求まるとする。

【０１３１】$ijnj>0$または$mjnj>0$のとき、

$$sim_{kj}=100-200\times|ijnj-mjn_k|/(ijnj+mjn_k)$$

$ijnj=0$かつ$mjn_k=0$のとき、

$sim_{kj}=-100$

上記動作をＩＴとＭＴ１を例として具体的に説明すると以下のようになる。th6 ＝75、it_1の接合部数＝22、it_2の接合部数＝６、mt_{11}の接合部数＝25、mt_{12}の接合部数＝６とすると式から、

mt_{11}とit_1の間の類似度：$sim_{11}=86.67$ となり、$sim_{11}>th6$ であるから、対応可能であるとして、その組（mt_{11}，it_1）を対応可能ペアとして保持する。

【０１３２】mt_{11}とit_2の間の類似度：$sim_{12}=-22.58$ となり、$sim_{12}<th6$ であるから、対応不可能であるとする。

【０１３３】mt_{12}とit_1の間の類似度：$sim_{21}=-14.28$ となり、$sim_{21}<th6$ であるから、対応可能であるとする。

【０１３４】mt_{12}とit_2の間の類似度：$sim_{22}=100$となり、$sim_{22}>th6$であるから、その組（mt_{12}，it_2）を対応可能ペアとして保持する。

【０１３５】以上の結果、対応可能ペアとして、ＩＴとＭＴ１との間では、p1＝（mt_{11}，it_1）とp2＝（mt_{12}，it_2）が検出された。

【０１３６】次に、異種対応可能ペア間両立関係判定部３５ｂ－２は、当該モデル文書と入力文書の間で、対応可能ペア検出部３５ｂ－１で検出された２つの異なる対応可能ペアが両立するものであるか否かを判定する。

【０１３７】ここでいう「両立する」ということは、２つの対応可能ペアが同時に存在することに矛盾が無いことを意味する。ここでの判定条件としては以下のものが上げられる。判定対象となる対応可能ペアをそれぞれ（mt_k，it_j）、（mt'_k，it'_j）とする。ここで、mt_kの左上端の位置座標を（$mt_k x1$，$mt_k y1$）、右下端の位置座標を（$mt_k x2$，$mt_k y2$）、mt'_kの左上端の位置座標を（$mt'_k x1$，$mt'_k y1$）、右下端の位置座標を（$mt'_k x2$，$mt'_k y2$）、it_jの左上端の位置座標を（$it_j x1$，$it_j y1$）、右下端の位置座標を（$it_j x2$，$it_j y2$）、it'_jの左上端の位置座標を（$it'_j x1$，$it'_j y1$）、右下端の位置座標を（$it'_j x2$，$it'_j y2$）とする。

【０１３８】また、文書画像の座標系（ｘ，ｙ）：０＜

ｘ≦ＷＩＤＴＨ，０＜ｙ≦ＨＥＩＧＨＴにおいて、図１８に示すような座標系を定義する。すなわち、図１８（ａ）中に示す矩形領域Ｔ：（左上端の位置座標を（ｔｘ１，ｔｙ１）、右下端の位置座標を（ｔｘ２，ｔｙ２）とする）に対して、図１８（ｂ）に示す０＜ｘ＜ｔｘ１かつ０＜ｙ≦ＨＥＩＧＨＴを満たす領域を領域１、図１８（ｃ）に示すｔｘ２＜ｘ≦ＷＩＤＴＨかつ０＜ｙ≦ＨＥＩＧＨＴを満たす領域を領域２、図１８（ｄ）に示す０＜ｘ≦ＷＩＤＴＨかつ０＜ｙ＜ｔｙ１を満たす領域を領域３、図１８（ｅ）に示す０＜ｘ≦ＷＩＤＴＨかつｔｙ２＜ｙ≦ＨＥＩＧＨＴを満たす領域を領域４と定義する。

【０１３９】また、判定条件として、以下の条件１から条件３までのすべての条件を満たさない場合のみ、２つの対応可能ペアが両立可能であると見なす。

【０１４０】条件１：$mt_k = mt'_k$、
条件２：$it_j = it'_j$、
条件３：配置関係に逆転があること。条件３は、以下の条件３－１～３－４の条件を満たす場合である。すなわち、

条件３－１：it_j（it'_j）に対して、it'_j（it_j）が領域１にあり、mt_k（mt'_k）に対して、mt'_k（mt_k）が領域２にある。

【０１４１】条件３－２：it_j（it'_j）に対して、it'_j（it_j）が領域２にあり、mt_k（mt'_k）に対して、mt'_k（mt_k）が領域１にある。

【０１４２】条件３－３：it_j（it'_j）に対して、it'_j（it_j）が領域３にあり、mt_k（mt'_k）に対して、mt'_k（mt_k）が領域４にある。

【０１４３】条件３－４：it_j（it'_j）に対して、it'_j（it_j）が領域４にあり、mt_k（mt'_k）に対して、mt'_k（mt_k）が領域３にある。

【０１４４】異種対応可能ペア間両立関係判定部３５ｂ－２の動作を、前述した対応可能ペアp1とp2を用いて具体的に説明する。ＩＴとＭＴ１の間における２つの異なる対応可能ペアの組として（p1，p2）がある。このペアは条件１から条件３までのすべてを満たさないので両立可能であると判断される。異種対応可能ペア間両立関係判定部３５ｂ－２は、両立可能と判断された対応可能ペアを、最良マッチング抽出部３５ｂ－３に出力する。

【０１４５】最良マッチング抽出部３５ｂ－３は、異種対応可能ペア間両立関係判定部３５ｂ－２において両立すると判定された対応可能ペアのうち、すべてが互いに両立可能な対応可能ペアの最大の集合を求めることにより、入力表特徴集合とモデル表特徴集合間の最良マッチングを抽出する。

【０１４６】最良マッチング抽出部３５ｂ－３は、さらに図１９に示すように、連合グラフ作成部３５ｂ－３ａと最大クリーク抽出部３５ｂ－３ｂにより構成されている。連合グラフ作成部３５ｂ－３ａは、異種対応可能ペア間両立関係判定部３５ｂ－２の判定結果を受け取り、その情報をもとに連合グラフを作成する。この連合グラフの節点は対応可能ペアを示し、節点間を結ぶ弧は２つの対応可能ペアが両立可能であることを示すものであり、異種対応可能ペア間両立関係判定部３５ｂ－２の判定結果を表す補助的なデータ構造である。

【０１４７】具体的には、対応可能ペアp1とp2が連合グラフ作成部３５ｂ－３ａに送られ、この結果、図２０に示す連合グラフが得られる。上述した「すべてが互いに両立可能な対応可能ペアの最大の集合」はこの連合グラフにおいては全体的に連結した（完全に互いに両立可能な）最大の節点集合であると捉えることができる。それはクリークであり、本実施例ではより大きなクリークはより良いマッチングを表している。

【０１４８】最大クリーク検出部３５ｂ－３ｂは、連合グラフ作成部３５ｂ－３ａによって作成された連合グラフから、節点数が最大である完全連結部分グラフ（最大クリーク）を抽出する。なお、最大クリーク検出部３５ｂ－３ｂの動作は、例えば文献「信学論（Ｄ），J68-D，3,pp221-228,1985」に記載されている方式を適用することができる。

【０１４９】最大クリーク検出部３５ｂ－３ｂによって抽出されたクリーク、すなわち、「すべてが互いに両立可能な対応可能ペアの最大の集合」を構成する要素の対応可能ペアは、入力文書の表特徴集合ＩＴと当該モデル文書の表特徴集合ＭＴ１との間で対応の取れた表の組を表す。具体的には、最大クリーク：mclique ＝p1，p2が得られ、$mt1_1$とit_1、$mt1_2$とit_2の対応がそれぞれ得らたことになる。この対応結果は、後段の罫線照合部３５ｃに送られる。

【０１５０】このとき、入力文書とモデル文書の表特徴集合の各表では、その内部に含まれる罫線特徴の座標値は、表の左上端の座標値で正規化されている。また、入力文書の各表の罫線特徴の座標値は、対応付くモデル文書の表に重ね合わせられるようにスケール変換されている。

【０１５１】すなわち、入力表の縦幅がｉｔ-height、横幅がｉｔ-width、対応付いたモデル表の縦幅がｍｔ-height、横幅がｍｔ-widthであるとき、当該入力表の正規化された罫線特徴の座標値は、さらに、ｘ座標値では（ｍｔ-width／ｉｔ-width）倍、ｙ座標値では（ｍｔ-height／ｉｔ-height）倍されることによりスケール変換される。この結果、後段の罫線照合部３５ｃでは同じスケールかつ同じ座標系のもとで入力表とモデル表に含まれる罫線集合間の対応付け処理が可能になる。

【０１５２】前述した対応可能ペア検出部３５ｂ－１の動作を示すＳｔｅｐ３において、ｓｉｍkj＜th6 である

場合には、ｍｔｋとｉｔｊの構造が異なっている他に、例えば図２１（ａ）に示すように、入力文書の印刷品質が悪いために表が分離している場合（図中（ａ－１））や、隣接する表の間隔が狭いために画像入力時に接触してしまったりする場合（図中（ｂ－１））も考えられる。

【０１５３】このような問題点に対応するために対応可能ペア検出部３５ｂ－１は、さらに以下の処理を実施するようにしてもよい。

【０１５４】Ｓｔｅｐ３－１：ｍｊｎｋ＞ｉｊｎｊ（ｍｊｎｋ＜ｉｊｎｊ）である場合には、ＩＴ（ＭＴ１）の中から以下の条件を満たすｉｔｊ（ｍｔｋ）に隣接する表特徴ｉｔｌ（ｍｔｌ）を１つ検出し、それらを統合して仮想的な表特徴ｉｔ′ｊを新たに生成し、その際に生じる接合部数ｉｊｎ′ｊ（ｍｊｎ′ｋ）とｍｊｎｋ（ｉｊｎｊ）との類似度ｓｉｍｌ′kjを計算する。

【０１５５】類似度：ｓｉｍｌ′kjがしきい値th6 以上である場合には、対応可能であるとして、（ｍｔｋ，ｉｔ′ｊ）（（ｍｔ′ｋ，ｉｔｊ））を対応可能ペアとして保持する。

【０１５６】条件：ｉｔｊ（ｍｔｋ）とｉｔｌ（ｍｔｌ）の間に他の表が存在しないこと。

【０１５７】Ｓｔｅｐ３－２：Ｓｔｅｐ３－１で求めたｓｉｍｌ′kjがしきい値th6 に満たない場合、類似度：ｓｉｍｌ′kjがしきい値th6 以上となるまで、あるいは条件を満たす統合すべき表が見つからなくなるまでＳｔｅｐ３－１を繰り返す。

【０１５８】この場合、表照合部３５ｂの最良マッチング部３５ｂ－３では、最大クリークを構成する「対応付き」のすべての組み合わせを出力するようにして、その後に次の条件を適用することにより候補を絞り込む。それでもなお複数の組み合わせが生じている場合には、それらすべてを後段の罫線照合部３５ｃに送り込み、その結果に基づいて（それらの中で照合度が最も高くなる組み合わせを選ぶことにより）最終的に表間対応関係を一意に決めるようにしてもよい。

【０１５９】条件：モデル表特徴集合のすべての要素が入力表特徴集合のいずれかの要素に対応していること。

【０１６０】次に、罫線照合部３５ｃでは、表照合部３５ｂで対応付いた入力表とモデル表にそれぞれ含まれる罫線集合間で対応付け処理を行なう。このとき水平罫線集合間の対応付けと垂直罫線集合間の対応付けを独立に行なうようにしてもよい。そして両者の整合を、各々の対応付け処理が済んだあとで得るようにしてもよい。この場合の罫線照合部３５ｃは、さらに図２２に示すように構成される。すなわち、表対応選択部３５ｃ－１、垂直罫線照合部３５ｃ－２、水平罫線照合部３５ｃ－３、及び方向間整合獲得部３５ｃ－４によって構成されている。

【０１６１】まず、表対応選択部３５ｃ－１は、表照合部３５ｂで抽出された表間対応付きの中から任意の対応を選択する。垂直罫線照合部３５ｃ－２は、表対応選択部３５ｃ－１によって選択された入力表とモデル表の垂直罫線集合間で対応付けを行なう。この対応付けに成功すれば、さらに水平罫線照合部３５ｃ－３は、水平罫線集合間で対応付けを行なう。方向間整合獲得部３５ｃ－３は、垂直罫線照合部３５ｃ－２による対応付けと水平罫線照合部３５ｃ－３による対応付けとの間の整合を獲得する。

【０１６２】なお、垂直罫線照合部３５ｃ－２と水平罫線照合部３５ｃ－２は、さらに図２３に示すように構成されている。これらの処理動作は、罫線方向の違いを考慮する以外は、基本的に同じである。

【０１６３】以下に、具体的な説明として、垂直罫線照合部３５ｃ－２における、ｍｔ１１の垂直罫線集合M_1 ＶＬ$_1$とｉｔ$_1$の垂直罫線集合ＩＶＬ$_1$の間の対応付け処理の動作について説明する。ただし、図１４より、M_1 ＶＬ$_1$＝（m_1 ｖｌ１，m_1 ｖｌ２，m_1 ｖｌ３，m_1 ｖｌ４，m_1 ｖｌ５，m_1 ｖｌ６，m_1 ｖｌ７，m_1 ｖｌ８）とする。以後、垂直罫線を単に罫線と略して説明する。

【０１６４】図２３中に示す対応可能罫線特徴ペア検出部３５ｃ－２ａは、モデルの罫線集合の各要素に対して、要素と対応付く可能性のある入力の罫線特徴をすべて検出し、対応可能な罫線特徴のペアとして管理する。すなわち、以下に説明する手順を実行する。

【０１６５】Ｓｔｅｐ１：任意のM_1 ＶＬ$_1$から任意の一つの罫線特徴m_1 ｖｌ$_k$（ｋ番目の罫線特徴を意味する）を選ぶ。

【０１６６】Ｓｔｅｐ２：m_1 ｖｌ$_k$に対応付く入力罫線特徴を検出するための探索範囲（ａｒｅａｍｖｌ）を設定する。

【０１６７】ここで、m_1 ｖｌ$_k$の罫線特徴の左上端と右下端の位置座標を（ｍｌｘ１，ｍｌｙ１）、（ｍｌｘ２，ｍｌｙ２）とすると、探索範囲は例えば（ｍｌｘ１－th9，ｍｌｙ１）と（ｍｌｘ２＋th9，ｍｌｙ２）の座標値で構成される矩形の内部としてもよい。ここで、しきい値th9 が予め与えられているものとする。

【０１６８】Ｓｔｅｐ３：Ｓｔｅｐ２で求められた探索範囲に内包・交差する入力罫線特徴を抽出する。

【０１６９】この処理は例えば以下のようにして行なわれる。すなわち、探索対象となる入力罫線の左上端と右下端の位置座標値を（ｉｘ１，ｉｙ１）、（ｉｘ２，ｉｙ２）とした場合、以下の条件を満たす入力罫線を抽出する。

【０１７０】条件：ｍｉｎ（ｍｌｘ２＋th9，ｉｘ２）－ｍａｘ（ｍｌｘ１＋th9，ｉｘ１）＋１＞０でかつｍｉｎ（ｍｌｙ２，ｉｙ２）－ｍａｘ（ｍｌｙ１，ｉｙ１）＋１＞０である。

【０１７１】Ｓｔｅｐ４：抽出された入力罫線特徴の一

つを選ぶ（例えば、ｉｖｌｊ（入力罫線特徴集合のｊ番目の罫線特徴）を選んだものとする）。

【０１７２】Ｓｔｅｐ５：m1 ｖｌｋの縦幅：ｍｌ-height とｉｖｌｊの縦幅：ｉｌ-height との類似度：ｓｉｍｌkjを計算する。

【０１７３】ここで、類似度：ｓｉｍｌkjは例えば次の式で求まるとする。

【０１７４】

ｍｌ-height ＞０またはｉｌ-height ＞０のとき

ｓｉｍｌkj＝ 100－ 200×｜ｍｌ-height －ｉｌ-height ｜／（ｍｌ-height＋ｉｌ-height ）

ｍｌ-height ＝０かつｉｌ-height ＝０のとき

ｓｉｍｌkj＝－100

Ｓｔｅｐ６：類似度：ｓｉｍｌkjがあらかじめ設定したしきい値th10以上である場合には、対応可能であるとして、その組（m1 ｖｌｋ，ｉｖｌｊ）を対応可能ペアとして保持する。

【０１７５】Ｓｔｅｐ７：類似度：ｓｉｍｌkjがしきい値th10未満である場合には、さらに以下の処理を行なう。

【０１７６】Ｓｔｅｐ７－１：図２４に示すように、１本であるべき線がかすれたり、途切れたりしているために分離している場合に対応するために以下の処理を行なう。

【０１７７】Ｓｔｅｐ７－１－１：Ｓｔｅｐ３で抽出された入力罫線特徴のうち、以下の条件を満たす罫線特徴のうち最も近接するものを１つ抽出し、それらを図２５に示すように統合した際に生じる縦幅：ｉｌ-height′とｍｌ-height との類似度：ｓｉｍｌ′kjを計算する。ただし、ｉｖｌｊの左上端と右下端の位置座標値を、（ｉｖｌｊｘ1，ｉｖｌｊｙ1）、（ｉｖｌｊｘ2，ｉｖｌｊｙ2）、抽出対象となる入力罫線の左上端と右下端の位置座標値を（ｉｘ1，ｉｙ1）、（ｉｘ2，ｉｙ2）とする。

【０１７８】条件：ｍｉｎ（ｉｖｌｊｘ2，ｉｘ2）－ｍａｘ（ｉｖｌｊｘ1，ｉｘ1）＋１＞０、類似度：ｓｉｍｌ′kjがしきい値th10以上である場合には、対応可能であるとして、その組（m1 ｖｌｋ，ｉｖｌ′ｊ）を対応可能ペアとして保持する。

【０１７９】Ｓｔｅｐ７－１－２：Ｓｔｅｐ７－１－１で求めたｓｉｍｌ′kj＜th10である場合、類似度：ｓｉｍｌ′kjがしきい値th10以上となるまで、あるいは条件を満たす統合すべき罫線が見つからなくなるまでＳｔｅｐ７－１－１を繰り返す。

【０１８０】Ｓｔｅｐ７－２：図２６に示すようにモデルの複数本の罫線特徴と入力の複数本の罫線特徴とが対応付くような場合に対応するために以下の処理を行なう。

【０１８１】Ｓｔｅｐ７－２－１：ｍｌ-height ＞ｉｌ-height ′（ｍｌ-height ＜ｉｌ-height ）である場合には、ＩＶＬ1（M1 ＶＬ1）の中から以下の条件を満たすｉｖｌｊ（m1 ｖｌｋ）に最も近接する罫線特徴ｉｖｌ′ｊ（m1 ｖｌ′ｋ）を１つ検出し、それらを統合した際に生じる縦幅ｉｌ-height ′（ｍｌ-height ′）とｍｌ-height(ｉｌ-height)との類似度：ｓｉｍｌ′kjを計算する。

【０１８２】ただし、ｉｖｌｊの左上端と右下端の位置座標値を、（ｉｖｌｊｘ1′，ｉｖｌｊｙ1′）、（ｉｖｌｊｘ2′，ｉｖｌｊｙ2′）、予め設定されたしきい値をth11とする。

【０１８３】条件：ｍｉｎ（ｉｖｌｊｘ２＋th11，ｉｖｌｊｘ2′）－ｍａｘ（ｉｖｌｊｘ１－th11，ｉｖｌｊｘ1′）＋１＞０

類似度：ｓｉｍｌ′kjがしきい値th10以上である場合には、対応可能であるとして、（m1 ｖｌｋ，ｉｖｌｊ）と（m1 ｖｌｋ，ｉｖｌ′ｊ）（（m1 ｖｌｋ，ｉｖｌｊ）と（m1 ｖｌ′ｋ，ｉｖｌｊ））を対応可能ペアとして保持する。

【０１８４】Ｓｔｅｐ７－２－２：Ｓｔｅｐ７－２－１で求めたｓｉｍｌ′kj＜th10である場合、類似度：ｓｉｍｌ′kjがしきい値th10以上となるまで、あるいは条件を満たす統合すべき罫線が見つからなくなるまでＳｔｅｐ７－２－１を繰り返す。

【０１８５】Ｓｔｅｐ８：Ｓｔｅｐ４～Ｓｔｅｐ７までをＳｔｅｐ３で抽出されたすべての入力罫線特徴に対して適用する。

【０１８６】Ｓｔｅｐ９：Ｓｔｅｐ１～Ｓｔｅｐ８までを着目M1 ＶＬ1 のすべての要素に対して適用する。

【０１８７】ここで、ｍｉｎ（）は（）内の２変数の内、小さい方を出力する関数であり、ｍａｘ（）は（）内の２変数の内、大きい方を出力する関数である。

【０１８８】対応可能罫線特徴ペア検出部３５ｃ－２ａの前述した処理動作を、M1 ＶＬ1とＩＶＬ1 を用いて具体的に説明する。対応可能罫線特徴ペア検出部３５ｃ－２ａにおいて、M1 ＶＬ1 の各要素で探索範囲を設けて、それぞれ対応付く可能性を持つＩＶＬ1 の要素を抽出した結果、以下のような罫線特徴ペアが得られたものとする。

【０１８９】ｐｌ1 ＝（m1 ｖｌ1，ｉｖｌ1）、
ｐｌ2 ＝（m1 ｖｌ2，ｉｖｌ2）、
ｐｌ3 ＝（m1 ｖｌ3，ｉｖｌ3）、
ｐｌ4 ＝（m1 ｖｌ4，ｉｖｌ2）、
ｐｌ5 ＝（m1 ｖｌ4，ｉｖｌ4）、
ｐｌ6 ＝（m1 ｖｌ5，ｉｖｌ4）、
ｐｌ7 ＝（m1 ｖｌ6，ｉｖｌ4）、
ｐｌ8 ＝（m1 ｖｌ7，ｉｖｌ5）、
ｐｌ9 ＝（m1 ｖｌ8，ｉｖｌ6）。

【０１９０】このうち、ｐｌ1、ｐｌ2、ｐｌ3、ｐｌ4、ｐｌ8、ｐｌ9は、Ｓｔｅｐ２からＳｔｅｐ６までの処理（以後、Ｓｔｅｐ２からＳｔｅｐ６までの処理で

得られた対応可能特徴ペアのみを１対１対応可能ペアと呼ぶ）で得られる。

【０１９１】ｐｌ5は、Ｓｔｅｐ２からＳｔｅｐ６までの処理で類似度：ｓｉｍｌ44がしきい値th10未満であったために、Ｓｔｅｐ７－２－１によりｉｖｌ4と「m1 ｖｌ4とm1 ｖｌ5を統合したもの」のペアを検出した。そして、そのペアをｐｌ5とｐｌ6の対応可能なペアに分けて管理している。ｐｌ7についても同様にＳｔｅｐ７－２－１でｉｖｌ4と「m1 ｖｌ4とm1 ｖｌ6を統合したもの」のペアを検出していることにより対応可能なペアとして抽出されている。

【０１９２】対応可能罫線特徴ペア検出部３５ｃ－２aにおける処理動作のＳｔｅｐ２における探索範囲の設定は、例えば以下に述べる手順で行なわれてもよい。例えば、図２７に示すモデル垂直罫線VL_1の探索範囲は、モデル垂直罫線VL_1に左側で隣接するモデル垂直罫線VL_2と、右側で隣接するモデル垂直罫線VL_3との距離に応じて設定するようにしてもよい。

【０１９３】すなわち、VL_1の左上端の位置座標値を（VL_1 x1，VL_1 y1）、右下端の位置座標値を（VL_1 x2，VL_1 y2）、VL_2の左上端の位置座標値を（VL_2 x1，VL_2 y1）、右下端の位置座標値を（VL_2 x2，VL_2 y2）、VL_3の左上端の位置座標値を（VL_3 x1，VL_3 y1）、右下端の位置座標値を（VL_3 x2，VL_3 y2）とすると、VL_1とVL_2の間の距離：ｄｉｓｔ12と、VL_1とVL_3の間の距離：ｄｉｓｔ13はそれぞれ、

ｄｉｓｔ12＝VL_1 x1 －VL_2 x2 ＋1、

ｄｉｓｔ13＝VL_3 x1 －VL_1 x2 ＋1、より求まることとする。

【０１９４】そして探索範囲を、（(VL_1 x1 －ｄｉｓｔ12／2），（VL_1 y1 ＋th9））と、（(VL_1 x2 ＋ｄｉｓｔ13／2），（VL_1 y2 ＋th9））、の位置座標値で構成される矩形領域としてもよい。

【０１９５】ここで、VL_1とその左側で隣接するモデル垂直罫線VL_2は、min(VL_1 y2，VL_2 y2）－max(VL_1 y1，VL_2 y1）＋1＞th13を満たす、距離ｄｉｓｔ12が最小であるモデル垂直罫線を検出することにより求めることができ、右側で隣接するモデル垂直罫線VL_3は、min(VL_1 y2，VL_3 y2）－max(VL_1 y1，VL_3 y1）＋1＞th13を満たす、距離ｄｉｓｔ13が最小であるモデル垂直罫線を検出することにより求めることができる。ここでth13をしきい値とする。各モデル水平罫線の探索範囲も同様に求めることができる。

【０１９６】この他にも、Ｓｔｅｐ２における探索範囲を次のようにして設定してもよい。例えばｋ番目のモデル罫線に着目したとき、探索対象となっている入力罫線のうちｋ±α以内の番号を有するものを着目モデル罫線の探索範囲とするようにしてもよい。また、ある大きさのパラメータでスケール変換がなされた状態で、着目モデル罫線と同じ長さを持つ全ての入力罫線を探索範囲としても良い。

【０１９７】次に、対応可能罫線特徴ペア間両立性判定部３５ｃ－２bは、対応可能罫線特徴ペア検出部３５ｃ－２aで検出されたすべての２つの異なる対応可能ペアの組み合わせにおいて、それらが両立するものであるか否かを判定する。

【０１９８】ここでの判定条件としては以下のものが上げられる。判定対象となる対応可能ペアをそれぞれｐ＝（ｍｌｋ，ｉｌｊ）、ｐ′＝（ｍｌ′ｋ，ｉｌ′ｊ）とする。

【０１９９】ここで、ｍｌｋの左上端の位置座標を（ｍｌｋ ｘ1，ｍｌｋ ｙ1）、右下端の位置座標を（ｍｌｋ ｘ2，ｍｌｋ ｙ2）、ｍｌ′ｋの左上端の位置座標を（ｍｌ′ｋ ｘ1，ｍｌ′ｋ ｙ1）、右下端の位置座標を（ｍｌ′ｋ ｘ2，ｍｌ′ｋ ｙ2）、ｉｌｊの左上端の位置座標を（ｉｌｊ ｘ1，ｉｌｊ ｙ1）、右下端の位置座標を（ｉｌｊ ｘ2，ｉｌｊ ｙ2）、ｉｌ′ｊの左上端の位置座標を（ｉｌ′ｊ ｘ1，ｉｌ′ｊ ｙ1）、右下端の位置座標を（ｉｌ′ｊ ｘ2，ｉｌ′ｊ ｙ2）とする。

【０２００】判定条件は、条件１から条件４までのすべての条件を満たさない場合のみ２つの対応可能ペアが両立可能であるとみなす。

【０２０１】条件１：ｐとｐ′のどちらかが１対１対応可能ペアでり、かつｍｌｋ＝ｍｌ′ｋである。

【０２０２】条件２：ｐとｐ′のどちらかが１対１対応可能ペアでり、かつｉｌｊ＝ｉｌ′ｊである。

【０２０３】条件３：ｐとｐ′のどちらも１対１対応可能ペアでり、かつ以下の条件３－１，３－２のどちらかを満たす。

【０２０４】条件３－１：（min(ｍｌｋ ｘ2，ｍｌ′ｋ ｘ2）－max(ｍｌｋ ｘ1，ｍｌ′ｋｘ1）＋1）＞0かつ（min(ｍｌｋ ｙ2，ｍｌ′ｋ ｙ2）－max(ｍｌｋ ｙ1，ｍｌ′ｋ ｙ1）＋1）＞0。

【０２０５】条件３－２：（min(ｉｌｋ ｘ2，ｉｌ′ｋ ｘ2）－max(ｉｌｋ ｘ1，ｉｌ′ｋｘ1）＋1）＞0かつ（min(ｉｌｋ ｙ2，ｉｌ′ｋ ｙ2）－max(ｉｌｋ ｙ1，ｉｌ′ｋ ｙ1）＋1）＞0。

【０２０６】条件４：配置関係に逆転がある。すなわち、以下の４－１～４－４の状態にある。

【０２０７】４－１：ｉｌｊ（ｉｌ′ｊ）に対して、ｉｌ′ｊ（ｉｌｊ）が領域１にあり、ｍｌｋ（ｍｌ′ｋ）に対して、ｍｌ′ｋ（ｍｌｋ）が領域２にある。

【０２０８】４－２：ｉｌｊ（ｉｌ′ｊ）に対して、ｉｌ′ｊ（ｉｌｊ）が領域２にあり、ｍｌｋ（ｍｌ′ｋ）に対して、ｍｌ′ｋ（ｍｌｋ）が領域１にある。

【０２０９】４－３：ｉｌｊ（ｉｌ′ｊ）に対して、ｉｌ′ｊ（ｉｌｊ）が領域３にあり、ｍｌｋ（ｍｌ′ｋ）に対して、ｍｌ′ｋ（ｍｌｋ）が領域４にある。

【０２１０】４－４：ｉｌｊ（ｉｌ′ｊ）に対して、ｉｌ′ｊ（ｉｌｊ）が領域４にあり、ｍｌｋ（ｍｌ′ｋ）に対して、ｍｌ′ｋ（ｍｌｋ）が領域３にある。

【０２１１】対応可能罫線特徴ペア間両立性判定部３５ｃ－２ｂの動作を、前述した対応可能罫線特徴ペアｐｌ１～ｐｌ９を用いて具体的に説明する。ｐｌ１からｐｌ９までの９個の対応可能罫線特徴ペアのすべての組み合わせは３６通りある。

【０２１２】そのうち上記条件１～４をすべて満たさない組み合わせは、（ｐｌ１，ｐｌ２）、（ｐｌ１，ｐｌ３）、（ｐｌ１，ｐｌ４）、（ｐｌ１，ｐｌ５）、（ｐｌ１，ｐｌ６）、（ｐｌ１，ｐｌ７）、（ｐｌ１，ｐｌ８）、（ｐｌ１，ｐｌ９）、（ｐｌ２，ｐｌ３）、（ｐｌ２，ｐｌ５）、（ｐｌ２，ｐｌ６）、（ｐｌ２，ｐｌ７）、（ｐｌ２，ｐｌ８）、（ｐｌ２，ｐｌ９）、（ｐｌ３，ｐｌ５）、（ｐｌ３，ｐｌ６）、（ｐｌ３，ｐｌ７）、（ｐｌ３，ｐｌ８）、（ｐｌ３，ｐｌ９）、（ｐｌ４，ｐｌ６）、（ｐｌ４，ｐｌ７）、（ｐｌ４，ｐｌ８）、（ｐｌ４，ｐｌ９）、（ｐｌ５，ｐｌ６）、（ｐｌ５，ｐｌ７）、（ｐｌ５，ｐｌ８）、（ｐｌ５，ｐｌ９）、（ｐｌ６，ｐｌ７）、（ｐｌ６，ｐｌ８）、（ｐｌ６，ｐｌ９）、（ｐｌ７，ｐｌ８）、（ｐｌ７，ｐｌ９）、（ｐｌ８，ｐｌ９）の３２通りである。これらの組み合わせの各々において、それを構成する対応可能罫線特徴ペアは、両立可能であると判断される。

【０２１３】次に、最良マッチング抽出部３５ｃ－２ｃは，対応可能罫線特徴ペア間両立性判定部３５ｃ－２ｂにおいて両立すると判定されたもののうち、すべてが互いに両立可能な対応可能ペアの最大の集合を求めることにより、入力表特徴集合とモデル表特徴集合間の最良マッチングを抽出する。

【０２１４】なお、最良マッチング抽出部３５ｃ－２ｃは、図１７に示す最良マッチング抽出部３５ｂ－３と同一の機能を有している（詳細は図１９に示している）。最良マッチング抽出部３５ｃ－２ｃを構成している連合グラフ作成部では、対応可能罫線特徴ペア間両立性判定部３５ｃ－２ｂで得られた結果から連合グラフを作るか、ｐｌ１からｐｌ９までの９個の対応可能罫線特徴ペアに対しては、図２８に示すものが作られる。

【０２１５】M_1 ＶＬ$_1$ とＩＶＬ$_1$ の間の対応付け処理に関しては、最良マッチング抽出部３５ｃ－２ｃにおいて、（ｍ$_1$ ｖｌ１，ｉｖｌ１）、（ｍ$_1$ ｖｌ２，ｉｖｌ２）、（ｍ$_1$ ｖｌ３，ｉｖｌ３）、（ｍ$_1$ ｖｌ５，ｉｖｌ４）、（ｍ$_1$ ｖｌ６，ｉｖｌ４）、（ｍ$_1$ ｖｌ７，ｉｖｌ５）、（ｍ$_1$ ｖｌ８，ｉｖｌ６）、の対応が抽出されたものとする。

【０２１６】ｍｔ１$_1$ の水平罫線集合：M_1 ＨＬ$_1$ ＝（ｍ$_1$ ｈｌ１，ｍ$_1$ ｈｌ２，ｍ$_1$ ｈｌ３，ｍ$_1$ ｈｌ４，ｍ$_1$ ｈｌ５）とｉｔ$_1$ の水平罫線集合：ＩＨＬ$_1$ ＝（ｉｈｌ１，ｉｈｌ２，ｉｈｌ３，ｉｈｌ４，ｉｈｌ５）の間の対応付け処理も、水平罫線照合部３５ｃ－３において、同様に、（ｍ$_1$ ｈｌ１，ｉｈｌ１）、（ｍ$_1$ ｈｌ２，ｉｈｌ２）、（ｍ$_1$ ｈｌ３，ｉｈｌ３）、（ｍ$_1$ ｈｌ４，ｉｈｌ４）、（ｍ$_1$ｈｌ５，ｉｈｌ５）、の対応が抽出されたものとする。

【０２１７】表照合部３５ｂで抽出されたｍｔ１$_2$ とｉｔ$_2$ の対応における、罫線集合間の対応付け処理も罫線照合部３５ｃで同様に行なわれ、ｍｔ１$_2$ の垂直罫線集合：M_1 ＶＬ$_2$ ＝（ｍ$_1$ ｖｌ９，ｍ$_1$ ｖｌ10，ｍ$_1$ ｖｌ11）とｉｔ$_2$ の垂直罫線集合：ＩＶＬ$_2$ ＝（ｉｖｌ７，ｉｖｌ８，ｉｖｌ９）の間では、（ｍ$_1$ ｖｌ９，ｉｖｌ７）、（ｍ$_1$ ｖｌ10，ｉｖｌ８）、（ｍ$_1$ ｖｌ11，ｉｖｌ９）、ｍｔ１$_2$ の水平罫線集合：M_1 ＨＬ$_1$ ＝（ｍ$_1$ ｈｌ６，ｍ$_1$ ｈｌ７）とｉｔ$_2$ の水平罫線集合：ＩＨＬ２＝（ｉｈｌ７，ｉｈｌ８）の間では、（ｍ$_1$ ｖｌ６，ｉｖｌ６）、（ｍ$_1$ ｖｌ７，ｉｖｌ７）の罫線特徴間の対応が得られたものとする。

【０２１８】照合度計算部３５ｄは、罫線照合部３５ｃによって対応関係が抽出されたモデル文書と入力文書の間で、当該構造化特徴間の対応付きを数量化することによりその度合い（照合度）を計算する。

【０２１９】照合度は、モデル照合部３５に送られてきたすべてのモデル文書と入力文書との間で計算され、照合結果出力部３５ｅに出力される。照合度：matching-metric は、モデル水平罫線数をsmhl-num、モデル垂直罫線数をsmvl-num、入力水平罫線のうちモデル水平罫線と対応付いたものの総数をsmch-num、入力垂直罫線のうちモデル垂直罫線と対応付いたものの総数をsmcv-numとしたときに、例えば以下の式で定義される。

【０２２０】matching-metric ＝ 100－ 200×（｜smhl-num－smch-num｜＋｜smvl-num－smcv-num｜）／（（smhl-num－smch-num）＋（smvl-num－smcv-num））

例えば、図９の入力文書と図１４のモデル文書との間の照合度： matching-metric1 は、smhl-num＝７，smvl-num＝11，smch-num＝７，smcv-num＝10より、

matching-metric$_1$ ＝ 100－ 200×（｜７－７｜＋｜11－10｜／（７＋７）＋（11＋11））＝94.44 、となる。

【０２２１】また、図９の入力文書と図１５のモデル文書との間の照合度： matching-metric$_2$ は、当該構造化特徴間の対応関係が抽出可能であったので（－100 ）を設定する。

【０２２２】照合結果出力部３５ｅは、モデル照合部３５に送られてきたすべてのモデル構造化特徴と入力構造化特徴との間の照合度のうち、最大値を示すモデル文書と入力文書の組み合わせを選び、それらの構造化特徴間の対応関係と共に照合結果判定部３７に出力する。

【０２２３】照合結果判定部３７は、モデル照合部３５で最大照合度を示したモデル文書と入力文書の構造化特徴間の対応関係（以後、構造化特徴間対応関係と呼ぶ）が獲得できたか否かを判定する。

【０２２４】すなわち、モデル照合部３５で計算された最大照合度が予め与えられているしきい値：th7 以上である場合には、構造化特徴間対応関係を獲得できたと判定する。また、最大照合度がth7 未満である場合には、構造化特徴間対応関係を獲得できなかったとして、入力文書を棄却して、次の文書の入力を実施する。

【０２２５】また、この場合、フォーマット種別同定部３３で選出されなかったモデルのすべてに対して、モデル照合部３５で照合処理を行ない、その結果をこの照合結果判定部３７で判定するようにしてもよい。こうすると、フォーマット種別同定部３３における処理誤りを救済でき、処理結果の精度が高まる。

【０２２６】具体的に説明すると、例えば、モデル照合部３５で図９の入力構造化特徴と図１４のモデル構造化特徴との間で最大照合度が計算されたとすると、その値：max-sim は matching-metric1 より、max-sim ＝94.44 であり、 th7＝60とするとth7<max-sim より、当該構造化特徴間の対応関係は獲得されたと見なされる。

【０２２７】構造化特徴間の対応関係は、例えば次のような形式により保持される。すなわち、入力罫線特徴モデル罫線特徴の各対応付きにおいて、入力罫線特徴のimageに、それに対応するモデル罫線特徴の識別子を格納し、モデル罫線特徴のimageに、それに対応付くモデル罫線の識別子を格納する。予め、それぞれの罫線集合の各罫線特徴のimage に－１をセットしておけば、対応付かない罫線特徴のimage には常に－１が設定されていることになる。

【０２２８】モデル照合部３５と照合結果判定部３７を経て獲得された構造化特徴対応関係のうちで最も重要なものは、入力文書の罫線集合とモデル文書の罫線集合の間の対応関係（以後、罫線特徴間対応関係と呼ぶ）である。後段の文字列領域抽出部４３は、この対応関係に基づいて入力画像から文字列領域を切り出す。

【０２２９】このとき罫線間対応関係が、どちらかの構造化特徴の欠落のため不完全であったり、矛盾を含んでいる場合には、処理不能となってしまう。このような問題点を解決するために、未対応・矛盾対応発見修正部４１は、文字列領域抽出部４３による処理の前に、罫線間対応関係に対する未対応・矛盾対応を修正する。

【０２３０】未対応・矛盾対応発見修正部４１は、以下の処理（１）（２）を実施する。

【０２３１】（１）モデルの罫線集合を構成する罫線特徴のうち入力罫線に対応付いてないものを検出し、すでに対応付いている他の対応関係を利用し、入力罫線集合において対応付くものを発見するか、対応付くべきものを仮想入力罫線として自動的に発生させて、新たに入力罫線集合に加える。

【０２３２】（２）上記（１）の処理を行なうと図２９に示すような矛盾が生じてしまう場合には、それを解消し、無矛盾な対応関係を生みだすようにする。

【０２３３】このような未対応・矛盾対応発見修正部３９の動作は、例えば図３０に示すフローチャートに従う。図３０に示すフローチャートの各ステップの処理動作は以下のようになる。ただし、この時点でも入力罫線集合の各要素は対応付いているモデル罫線集合の座標系に変換されたままであるものとする。

【０２３４】ｆ１：当該モデル罫線集合の要素のうち、特徴のimage に－１が付与されているものを検出する。

【０２３５】ｆ２：着目モデル罫線特徴の左上端の座標値（ｍｘ１，ｍｙ１）と右下端の座標値（ｍｘ２，ｍｙ２）に対して、それぞれ以下のように探索範囲を設ける。ここで、しきい値th8 が予め設定されているものとする。

【０２３６】
Ｉｉｍ-x1 ＝ｍｘ１ －th8
Ｉｉｍ-y1 ＝ｍｙ１ －th8
Ｉｉｍ-x2 ＝ｍｘ２ ＋th8
Ｉｉｍ-y2 ＝ｍｙ２ ＋th8

この探索範囲に含まれる入力罫線特徴のうち、着目モデル罫線に最も近いものを検出する。この時、近さの尺度を表す距離値は、例えば罫線特徴の重心間のユークリッド距離で定義されてもよい。

【０２３７】ｆ３：仮想入力罫線として、以下の特徴（罫線特徴）を有するものを生成する。罫線特徴は、左上端の位置座標（ｋｘ１，ｋｙ１）、右下端の位置座標（ｋｘ２，ｋｙ２）、縦幅（ｋ－height）、横幅（ｋ－width）、重心の位置座標（ｋｃｘ，ｋｃｙ）を含む。

【０２３８】ｆ４：着目モデル罫線が、

１．水平罫線の場合、着目モデル罫線の両端に接続する垂直モデル罫線をそれぞれ検出する。これらの垂直モデル罫線が、(a) 存在し、かつそれに対応付く入力罫線が存在する場合には、その垂直入力罫線特徴の左上端と右下端の位置座標のうちｘ座標値のみを、それぞれｋｘ１とｋｘ２に格納する。

【０２３９】(b) 存在しない場合、もしくは垂直モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目水平モデル罫線の位置座標値であるｍｘ１とｍｘ２をそれぞれｋｘ１とｋｘ２に格納する。

【０２４０】２．垂直罫線の場合、着目モデル罫線の位置座標値であるｍｘ１とｍｘ２をそれぞれｋｘ１とｋｘ２に格納する。

【０２４１】ｆ５：着目モデル罫線が、

１．垂直罫線の場合、着目モデル罫線の両端に接続する水平モデル罫線をそれぞれ検出する。これらの水平モデル罫線が、(a) 存在し、かつそれに対応付く入力罫線が存在する場合には、その垂直入力罫線特徴の左上端と右下端の位置座標のうちｙ座標値のみを、それぞれｋｙ１とｋｙ２に格納する。

【０２４２】(b) 存在しない場合、もしくは水平モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目垂直モデル罫線の位置座標値であるｍｙ

1とmy2 をそれぞれｋｙ1 とｋｙ2 に格納する。
【０２４３】２．水平罫線の場合、着目モデル罫線の位置座標値であるｍｙ1 とｍｙ2 をそれぞれｋｙ1 とｋｙ2 に格納する。
【０２４４】ｆ６：ｆ２で検出モデル罫線特徴のimageを調べる。image に－１がセットされている場合には、図３０のフローチャートでＮＯの方向に、－１以外の値がセットされている場合にはＹＥＳの方向に処理を進める。
【０２４５】ｆ７：着目モデル罫線が水平罫線の場合、着目モデル罫線の両端に接続する垂直モデル罫線をそれぞれ検出する。これらの垂直モデル罫線が、
１．存在し、かつそれら対応付く入力罫線が存在する場合には、その垂直入力罫線特徴の左上端と右下端の位置座標のうちｘ座標値のみをそれぞれ着目水平入力罫線の左上端と右下端のｘ座標値に格納する。
【０２４６】２．存在しない場合、もしくは垂直モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目水平モデル罫線の位置座標値であるｍｘ1とｍｘ2 をそれぞれ着目水平入力罫線の左上端と右下端のｘ座標値に格納する。
【０２４７】ｆ８：着目モデル罫線が垂直罫線の場合、着目モデル罫線の両端に接続する水平モデル罫線をそれぞれ検出する。これらの水平モデル罫線が、
１．存在し、かつそれに対応付く入力罫線が存在する場合には、その水平入力罫線特徴の左上端と右下端の位置座標のうちｙ座標値のみをそれぞれ着目水平入力罫線の左上端と右下端のｙ座標値に格納する。
【０２４８】２．存在しない場合、もしくは水平モデル罫線が存在してもそれに対応付く入力罫線が存在しない場合には、着目垂直モデル罫線の位置座標値であるｍｙ1とｍｙ2 をそれぞれ着目垂直入力罫線の左上端と右下端のｙ座標値に格納する。
【０２４９】図３０に示すフローチャートの具体的な動作を図９に示す入力構造化特徴と図１４に示すモデル構造化特徴を用いて説明する。当該モデル罫線特徴集合のうちm1 ｖ14 のみ、対応付く入力罫線が存在しない。従って、図３０のフローチャートのステップｆ１でm1 ｖ14 が検出される。次いで、ステップｆ２において、m1 ｖ14 の探索範囲に含まれるｉｖ14 を検出し、さらにステップｆ６でｉｖ14 に対応付くモデル罫線としてm1 ｖ15 とm1 ｖ16 を検出する。このうち、m1 ｖ16 とm1 ｖ14 は共存できる（同時に存在できる）が、m1 ｖ14とm1 ｖ15 の両立性は矛盾する。そこで、ステップｆ９でm1 ｖ14 とm1 ｖ15 のどちらが当該入力罫線に近いかが判断され、その結果、m1 ｖ14 の方が近いことが分かる。これによりステップｆ１０でｉｖ14 とm1 ｖ15 の対応関係が無効とされ、ステップｆ１１，ｆ１２において、m1 ｖ14 とｉｖ14 の対応が新たに生成され、これにより生じる座標値の変更がステップｆ７，ｆ８で行なわれる。
【０２５０】ステップｆ１０で対応関係を無効とされたことにより、m1 ｖ15 に対応付く入力罫線が存在しなくなった。これをステップｆ１４で検知したあと、m1 ｖ15 に対応付く入力罫線の発見・設定処理が行なわれる。まず、ステップｆ２でm1 ｖ15 の探索範囲でｉｖ14 を発見し、ステップｆ６でｉｖ14 に対応付くモデル罫線としてm1 ｖ14 とm1 ｖ16 を検出する。このうち、m1 ｖ16 とm1 ｖ15 は共存できる（同時に存在できる）が、m1 ｖ14 とm1 ｖ15 の両立性は矛盾することがわかっており、さらにm1 ｖ14 の方がm1 ｖ15 より近いことが分かっているので、m1 ｖ15 に対応する入力罫線は当該入力罫線集合では見つからなかったとして、ステップｆ３で仮想入力罫線を生成させ、その位置座標をステップｆ４とｆ５で設定し、ステップｆ１３で当該罫線集合に加える。この結果、当該モデル罫線集合に未対応モデル罫線がなくなり、未対応・矛盾対応発見修正手段３９における処理は終了となる。
【０２５１】以上の処理結果により得られた罫線マッチングに対して、以下の手順によって示されるマッチング後処理を適用することによって、さらにマッチング結果の精度を上げることができる。あるモデル罫線に対応付くべき入力罫線の付近に当該モデル罫線との間の類似度が高い線分（例えば取り消し線）が存在する場合、誤ってそれに対応づけてしまうことがある。以下の処理は、このような誤りを解消するものである。
【０２５２】Ｓｔｅｐ１：モデル罫線（ｍｌ）を１つ抽出する。
【０２５３】Ｓｔｅｐ２：その探索範囲（上述したａｒｅａml）内に存在する入力罫線（ｉｌ）を抽出し、モデルとの類似度（ｌｓｉｍml）を計算する。ここで、ｍｌｈ：ｍｌの縦幅、ｍｌｗ：ｍｌの横幅、ｉｌｈ：ｉｌの縦幅、ｉｌｗ：ｉｌの横幅とすると、ｍｌが水平罫線の場合には、
ｌｓｉｍml＝ 100－ 200×｜ｍｌｗ－ｉｌｗ｜／（ｍｌｗ＋ｉｌｗ）とし、垂直罫線の場合には、
ｌｓｉｍml＝ 100－ 200×｜ｍｌｈ－ｉｌｈ｜／（ｍｌｈ＋ｉｌｈ）とする。
【０２５４】Ｓｔｅｐ３：ｌｓｉｍml≧th10である入力罫線（ｉｌ′）をすべて抽出する。
【０２５５】Ｓｔｅｐ４：当該モデル罫線と対応づいていた入力罫線とｉｌ′の中で最も距離ｄｄmlが近い罫線を選び、あらためてそれを当該モデルに対応づける。ここで、ｍｌの左上端と右下端の座標をそれぞれ（ｍｌｘ１，ｍｌｙ１）、（ｍｌｘ2，ｍｌｙ２）、ｉｌ′の左上端と右下端をそれぞれ（ｉｌ′ｘ１，ｉｌ′ｙ１）、（ｉｌ′ｘ２，ｉｌ′ｙ２）とすると、ｍｌが水平罫線の場合には、
ｄｄml＝ min（ｍｌｙ１，ｉｌ′ｙ１）－ max（ｍｌｙ２，ｉｌ′ｙ２）＋１

とし、垂直罫線の場合には、
$dd_{ml} = \min(ml_{x1}, il'_{x1}) - \max(ml_{x2}, il'_{x2}) + 1$
とする。
【０２５６】以上の結果、入力構造化とモデル構造化特徴間の対応付け処理は終了したことになり、入力文書のすべての罫線特徴の座標系を元に戻すようにする。
【０２５７】未対応・矛盾対応発見修正手段３９により整合の得られた入力構造化特徴とモデル構造化特徴の対応関係を用いて、入力文書の罫線特徴集合に対して、以下に述べる罫線成形処理を施すことによって以後の処理が安定に行なわれるようにしてもよい。
【０２５８】例えば、図３１に示すように、罫線特徴（図中Ｌ１）の端点（図中Ｅ）がそれに直交する他の罫線特徴（図中Ｌ２）に接していない場合には、Ｌ１の端点の位置座標を変更して、Ｌ２に接するようにすることで罫線特徴を成形する。この成形処理は、入力画像の水平罫線集合と垂直罫線集合の両方に適用される。
【０２５９】水平罫線集合に対する罫線成形処理部の動作は、例えば次のようになる。ここでＬ１の左上端の座標値を（L_1 x1，L_1 y1）、右下端の座標値を（L_1 x2，L_1 y2）、Ｌ２の左上端の座標値を（L_2 x1，L_2 y1）、右下端の座標値を（L_2 x2，L_2 y2）とする。
【０２６０】Ｓｔｅｐ１：入力画像に対応するモデル文書の水平罫線特徴集合の任意の罫線特徴を選択する。
【０２６１】Ｓｔｅｐ２：着目水平罫線の両端に接するモデル垂直罫線を検出する。
【０２６２】例えば、図３１のＬ１の左端（点Ｅ）に接するべき垂直罫線Ｌ２は、次のようにして検出される。まず、探索領域を設定する。予め設定してあるしきい値をth12とすると、探索領域は例えば、（L_1 x1 －th12，（L_1 y1 ＋L_1 y2）／２－th12）、（L_1 x1 ＋th12，（L_1 y1 ＋L_1 y2）／２＋th12）で示される矩形（図３１中の破線で示されている矩形）の内部とする。
【０２６３】次に、探索領域に交差する垂直罫線のうち以下に定義する距離：ｄｉｓｔが最小のものを抽出する。
【０２６４】
$dist = |L_1 x1 - (L_2 x1 + L_2 x2)/2|$
Ｓｔｅｐ３：着目モデル水平罫線に対応付いている入力垂直罫線を抽出する。ここで、着目入力水平罫線の左上端の位置座標値を（ｉｘ１，ｉｙ１）、右下端の位置座標値を（ｉｘ２，ｉｙ２）とする。
【０２６５】Ｓｔｅｐ４：着目モデル水平罫線の左端に接するモデル垂直罫線に対応付いている入力垂直罫線を抽出する。ここで、その左上端の位置座標値を（ｌｘ１，ｌｙ１）、右下端の位置座標値を（ｌｘ２，ｌｙ２）とする。
【０２６６】Ｓｔｅｐ５：着目モデル水平罫線の右端に接するモデル垂直罫線に対応付いている入力垂直罫線を抽出する。ここで、その左上端の位置座標値を（ｒｘ１，ｒｙ１）、右下端の位置座標値を（ｒｘ２，ｒｙ２）とする。
【０２６７】Ｓｔｅｐ６：着目入力水平罫線の左上端の位置座標を（（ｌｘ１＋ｌｘ２）／２，ｉｙ１）に変更する。
【０２６８】Ｓｔｅｐ７：着目入力水平罫線の右下端の位置座標を（（ｒｘ１＋ｒｘ２）／２，ｉｙ２）に変更する。
【０２６９】Ｓｔｅｐ８：Ｓｔｅｐ２からＳｔｅｐ７までの処理を、すべてのモデル水平罫線特徴に対して行なう。
【０２７０】文書構造獲得部４１は、入力文書とモデル文書との間で矛盾のない構造化特徴間対応関係が得られた場合には、以下に示す処理を行なうことにより入力文書の構造を獲得する。すなわち、モデル登録部３１で予め登録されている当該モデル文書の知識を、構造化特徴間対応関係に基づいて入力文書にコピーすることにより、入力文書の構造を獲得したものと見なす。
【０２７１】具体的に説明すると、まず、モデル文書の各罫線特徴に付与されている識別番号（以後、ｉｄと呼ぶ）を罫線対応関係に基づいて、それに対応付いている入力文書の罫線特徴に付与する。
【０２７２】次いで、モデル文書に対して定義されている種々の知識を入力文書に対して用意されている、知識を格納するためのメモリ４３ｂの所定の領域にコピーする。後段の文字列領域抽出部４３、文字認識部４５、文字認識結果出力部４７は、メモリ４３ｂの所定の領域にコピーされた知識に基づいて動作する。
【０２７３】文字列領域抽出部４３は、文書構造獲得部４１で得られた当該文書に関する知識を用いて、認識対象文字列領域として定義されている領域のみを入力画像から切り出す。認識対象文字列領域は、例えば領域を囲んでいる上下左右の罫線のそれぞれの識別番号で定義されていてもよい。この場合、文字列領域抽出部４３では、入力文書画像中に位置するそれらの罫線の内側の部分を入力画像から切り出すことにより文字列領域を抽出する。
【０２７４】文字認識部４５は、文字列領域抽出部４３において抽出された文字列画像を、その文字列領域について定義されている知識を制約条件として用いて、例えば文献「信学技報、ＰＲＵ９３－４７、１９９３」に記載されている方式に基づいた処理により、文字切り出し／認識処理を行ない、コードデータに変換する。
【０２７５】このとき、各文字認識結果は類似度の降順にソートされており、上位Ｎ位まで保持されているようにしてもよい。また、認識結果はオペレータとシステムとの対話的な修正作業により修正されるようになってい

てもよい。

【０２７６】文字認識結果出力部４７は、文字認識部４５による文字認識結果に対して、文字単位で修正が済んだ文字コードデータに応じて、入力文書に対応付いたモデルに関して予め指定されている出力形態に基づき、ディスプレイあるいはファイルに出力する。

【０２７７】このようにして、本発明では、入力画像から抽出した罫線特徴を、特徴構造化部２９において構造化して、さらに表特徴、接合部に関する特徴などを抽出し、それらの関係を抽出・管理する。これらの情報を用いて、モデル登録部３１によって予め登録されているモデル文書のフォーマットに関する構造化特徴間で対応付け処理を行なう。このとき、モデル照合部３５によって、入力文書とモデル文書の間で表特徴集合間の照合処理を行ない、さらにその中に含まれる罫線特徴集合間で照合処理を行なうことにより、以下のような効果が得られる。

【０２７８】１．階層的な照合処理を行なうので計算量を少なくすることができる。

【０２７９】２．表が複数混在している場合も取り扱うことができる。

【０２８０】３．表単位の照合処理を行なうことにより全体的な配置関係を考慮することができ、局所的に見て同じ特徴量を有する場合でも対応付け誤りが生じない。

【０２８１】４．表間対応ごとに大きさの倍率に関するパラメータを求めることができ、モデル表に対応する入力表の大きさに関するパラメータがそれぞれ独立した値を持つような文書を取り扱うことができる。

【０２８２】５．表が印刷品質の劣化などにより分裂している場合も取り扱うことができる。

【０２８３】６．入力文書とモデル文書の間で罫線間の対応関係を求めるので、罫線がかすれていたり、途切れていたり、欠落している場合や余分な特徴抽出結果がある場合にも対応できる。

【０２８４】７．罫線単位で対応付け処理を行なうときに、複数対複数の対応を許しているので、罫線分布が局所的に変動している場合にも対応できる。

【０２８５】照合処理結果に対しては、照合結果判定部３７によって、照合度を用いてその妥当性を評価することにより、正しい対応付け結果のみを採用することができる。

【０２８６】さらに、照合処理結果に対して、未対応・矛盾対応発見修正部３９によって、不完全な対応結果を発見し、修正するので以下のような効果が得られる。

【０２８７】１．印刷の品質の悪い文書にも対応できる。

【０２８８】２．特徴抽出結果が不完全である場合にも対応できる。

【０２８９】３．対応関係に基づいている後段の処理で処理不能となることがない。

【０２９０】４．未対応箇所に対して、特徴抽出時のパラメータを調整して未検出な画像特徴の抽出が可能となる。

【０２９１】また、本発明ではモデル照合処理の前に、フォーマット種別同定部３３によって、入力文書の書式構造種別の同定処理を行なうことにより、照合処理で適用すべきモデルの種類を絞りこむので、無駄な照合処理を行なわないため、以下のような効果がある。

【０２９２】１．計算量が少ない。

【０２９３】２．構造のかけ離れたものにむりやり対応付けることがないため高精度な処理結果が得られる。

【０２９４】３．オペレータが対象文書ごとにモデルを手動で与える必要がないので、システムの自動運転が可能となる。

【０２９５】４．モデル登録時に正立したモデルフォーマットから、左および右に９０度回転させたもの、１８０度回転させたものの４種類を登録し、モデル照合時にこれらと対応付け処理を行なうことにより、文書の入力方向を限定しなくてもよい。

【０２９６】

【発明の効果】以上詳述したように本発明によれば、表形式の帳票などの書式構造を正確に認識でき、効率の良い文字列の領域の特定が可能となるものである。

【図面の簡単な説明】

【図１】本発明の一実施例に係わる画像処理装置の概略構成を示すブロック図。

【図２】本実施例における処理対象文書の一例を示す図。

【図３】本発明の一実施例である文字認識装置と組み合わせた文書画像処理システムの概略構成を示すブロック図。

【図４】本実施例における特徴抽出部２７の構成を示すブロック図。

【図５】入力画像から抽出された線分素の例を示す図。

【図６】入力画像から抽出された文字候補矩形の例を示す図。

【図７】本実施例における特徴抽出部２７の他の構成を示すブロック図。

【図８】文字候補矩形に交差・内包する線分素の例を示す図。

【図９】入力画像から抽出された罫線特徴の例を示す図。

【図１０】本実施例における特徴構造化部２９の構成を示すブロック図。

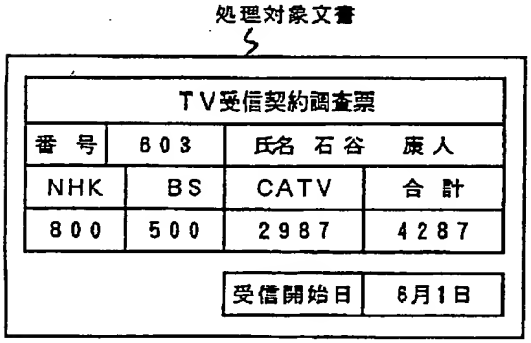
【図１１】入力画像から抽出された表特徴の例を示す図。

【図１２】入力画像から抽出された接合部特徴の例を示す図。

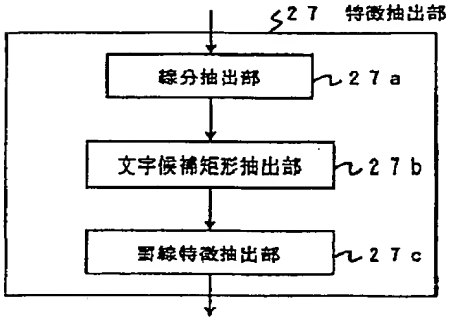
【図１３】階層的に関連づけられて管理される特徴に関する情報の一例を示す図。

【図１４】モデル文書の一例を示す図。
【図１５】モデル文書の一例を示す図。
【図１６】本実施例におけるモデル照合部３５の構成を示すブロック図。
【図１７】本実施例における表照合部３５ｂの構成を示すブロック図。
【図１８】任意の矩形領域の周辺に関する領域を定義するための説明に用いる図。
【図１９】本実施例における最良マッチング抽出部３５ｂ－３の構成を示すブロック図。
【図２０】連合グラフの一例を示す図。
【図２１】表特徴の分裂および接触の例を示す図。
【図２２】本実施例における罫線照合部３５ｃの構成を示すブロック図。
【図２３】本実施例における垂直罫線照合部３５ｃ－２および水平罫線照合部３５ｃ－３の構成を示すブロック図。
【図２４】１本の罫線が分離している場合の例を示す図。
【図２５】分離している罫線を統合した例を示す図。
【図２６】複数のモデル罫線と複数の入力罫線が対応付く例を示す図。
【図２７】任意の１本のモデル罫線特徴に対応付く可能性のある入力罫線特徴を抽出するための探索範囲の例を示す図。
【図２８】連合グラフの一例を示す図。
【図２９】未対応・矛盾対応発見抽出処理により矛盾した対応が生じてしまう例を示す図。
【図３０】未対応・矛盾対応発見抽出処理の流れ示すフローチャートの例を表す図。
【図３１】罫線特徴の端点がそれに直交する他の罫線特徴に接していない例を示す図。
【符号の説明】
１１，２１…画像入力部、１２…特徴抽出部、１３…特徴構造化部、１４…書式構造情報登録部、１５…書式構造種別同定部、１６…書式構造情報照合部、１７…未対応・矛盾対応発見修正部、１８…照合結果判定部、１９…文書構造獲得部、２３…２値化処理部、２５…前処理部、２７…特徴抽出部、２７ａ…線分抽出部、２７ｂ…文字候補矩形抽出部、２７ｃ…罫線特徴抽出部、２７ｄ…フィルタリング処理部、２９…特徴構造化部、２９ａ…罫線グループ化処理部、２９ｂ…表特徴抽出部、２９ｃ…罫線接合部検出部、２９ｄ…特徴間関係記述部、３１…モデル登録部、３３…フォント種別同定部、３５…モデル照合部、３５ａ…選択部、３５ｂ…表照合部、３５ｂ－１…対応可能ペア検出部、３５ｂ－２…異種対応可能ペア間両立関係判定部、３５ｂ－３…最良マッチング抽出部、３５ｂ－３ａ…連合グラフ作成部，３５ｂ－３ｂ…最大クリーク抽出部、３５ｃ…罫線照合部、３５ｃ－１…表対応選択部、３５ｃ－２…垂直罫線照合部、３５ｃ－２ａ…対応可能罫線特徴ペア検出部、３５ｃ－２ｂ…対応可能罫線特徴ペア間両立性判定部、３５ｃ－２ｃ…最良マッチング抽出部、３５ｃ－３…水平罫線照合部、３５ｃ－４…方向間整合獲得部、３５ｄ…照合度計算部、３５ｅ…照合結果出力部、３７…照合結果判定部、３９…未対応矛盾対応発見修正部、４１…文書構造獲得部、４３…文字列領域抽出部、４５…文字認識部、４７…文字認識結果出力部。

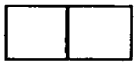
【図２】



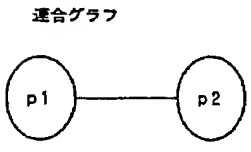
【図４】



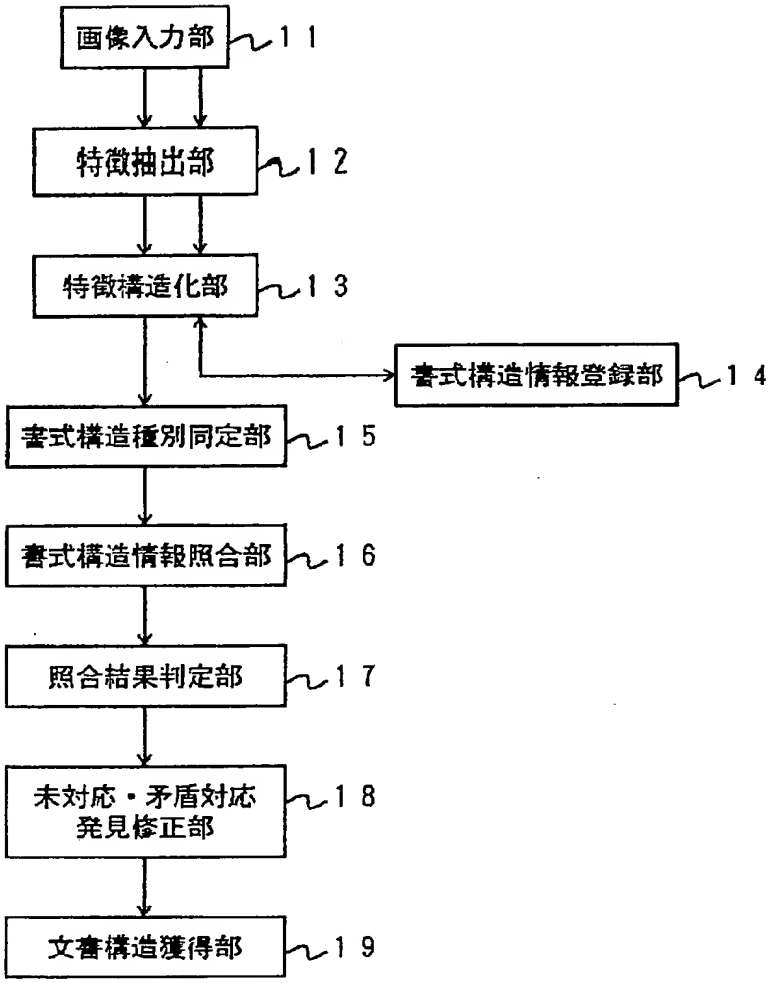
【図８】



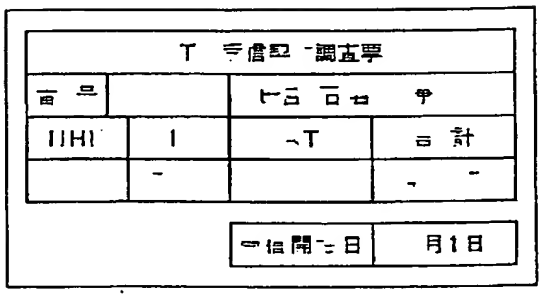
【図２０】



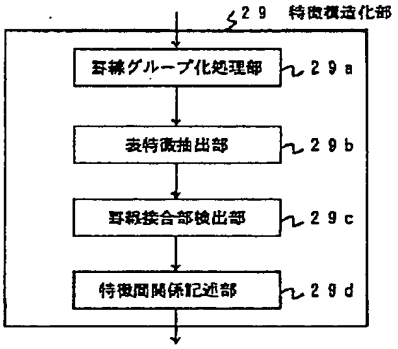
【図１】



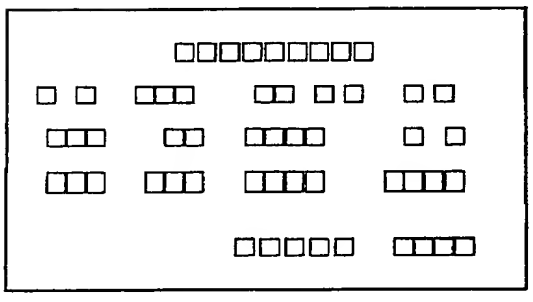
【図５】



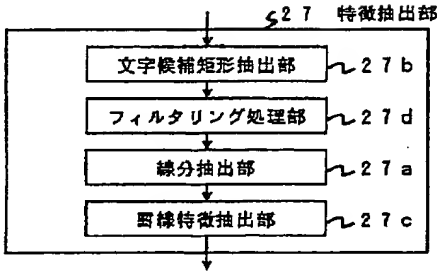
【図１０】



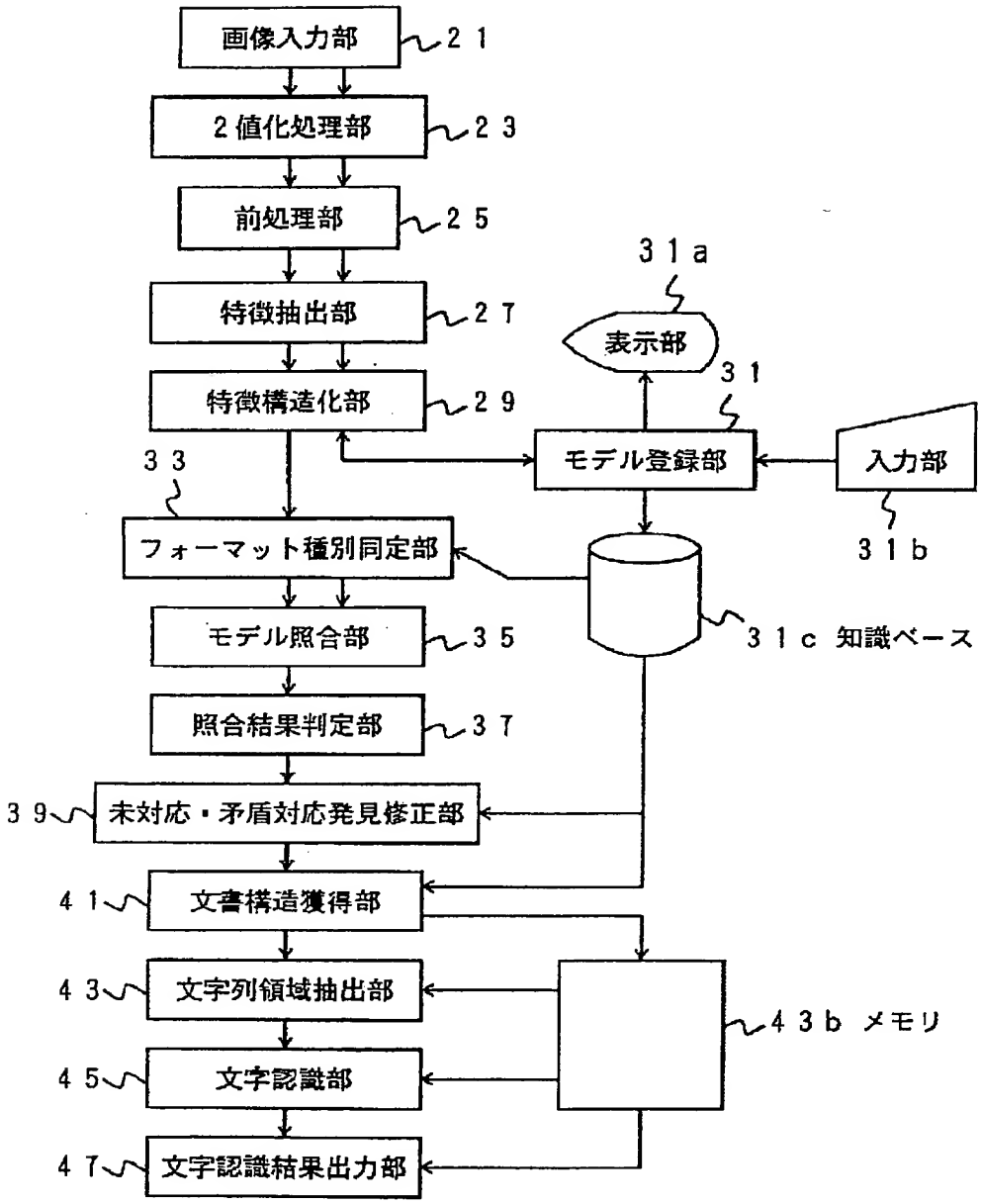
【図６】



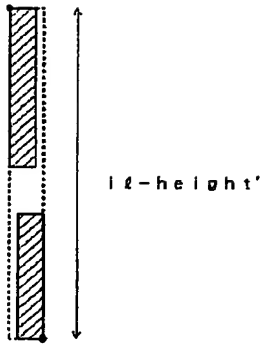
【図７】



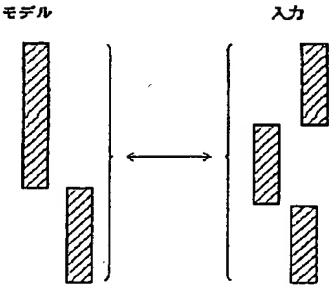
【図３】



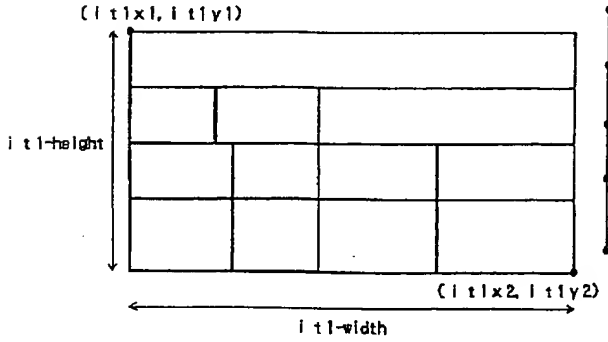
【図２５】



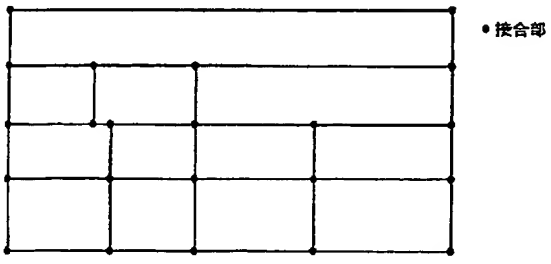
【図２６】



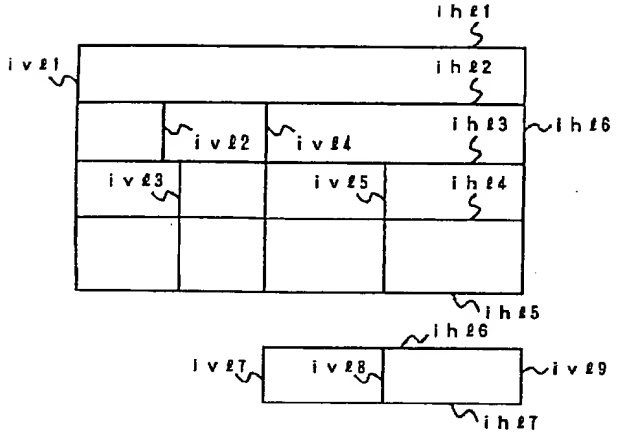
【図１１】



【図１２】

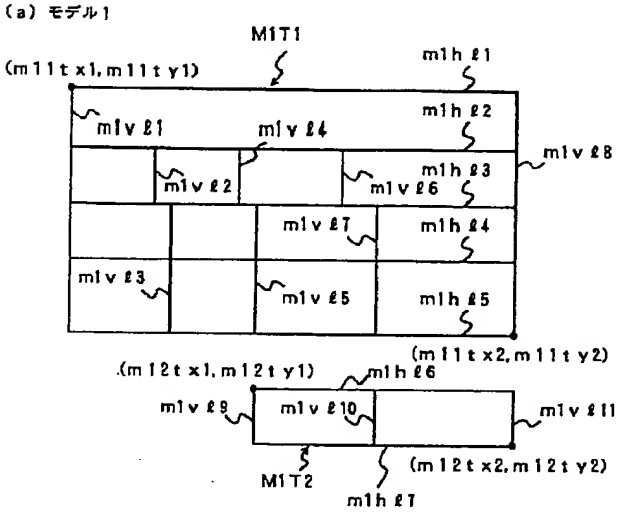


【図９】



【図１４】

(a) モデル1

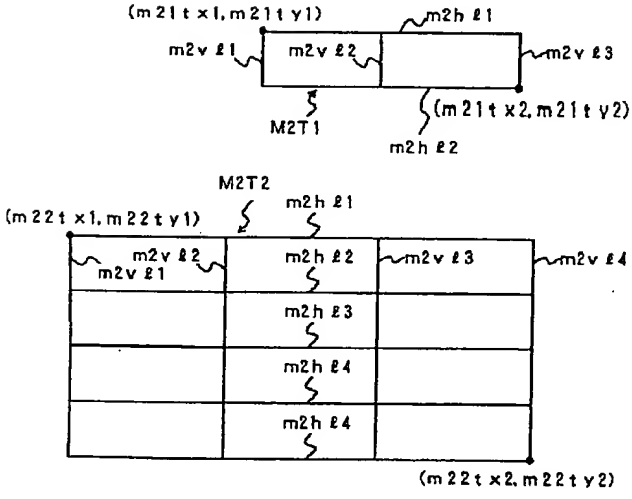


(b) ページ特徴

表　　数	2
表特徴集合	M1T1 = (M1T1 , M1T2)
接合部数	31
水平罫線数	7
垂直罫線数	11

【図１５】

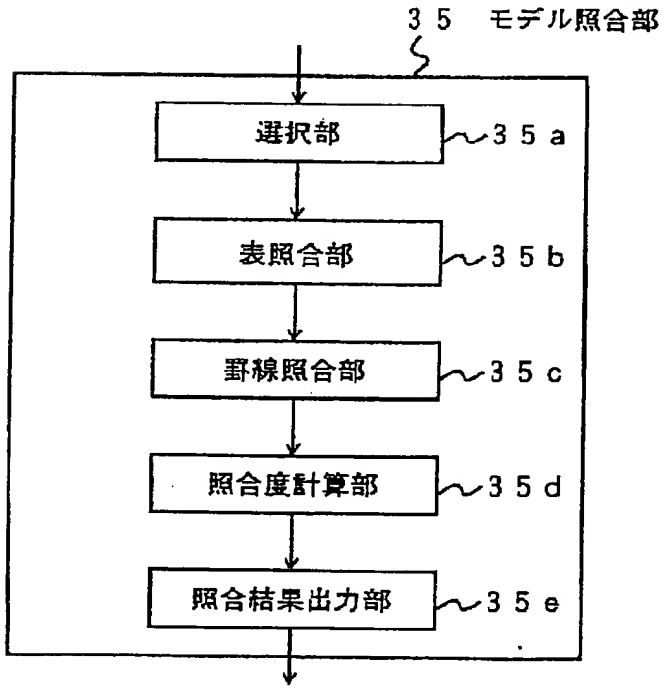
(a) モデル2



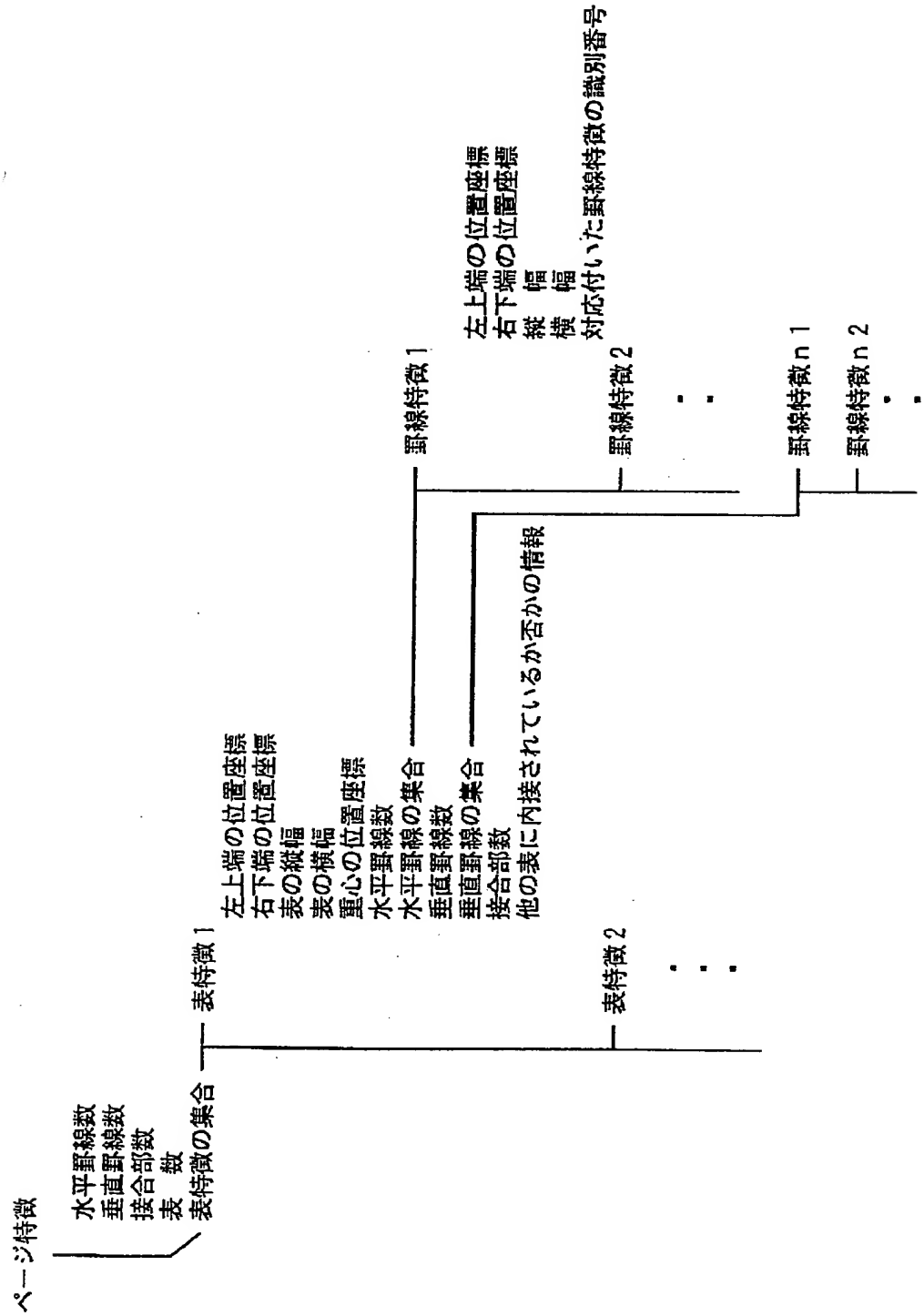
(b) ページ特徴

表　　数	2
表特徴集合	MT2= (M2T1 , M2T2)
接合部数	26
水平罫線数	7
垂直罫線数	7

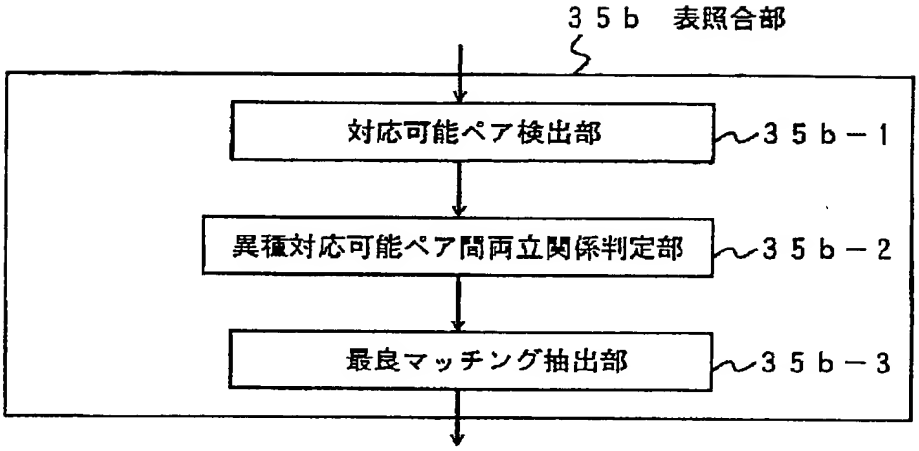
【図１６】



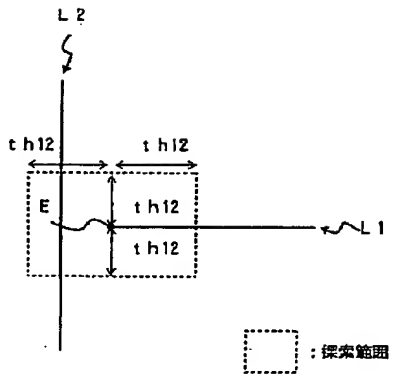
【図１３】


ページ特徴
水平罫線数
垂直罫線数
接合部数
表　数
表特徴の集合
表特徴１
左上端の位置座標
右下端の位置座標
表の縦幅
表の横幅
重心の位置座標
水平罫線数
水平罫線の集合
垂直罫線数
垂直罫線の集合
接合部数
他の表に内接されているか否かの情報
罫線特徴１
左上端の位置座標
右下端の位置座標
縦　幅
横　幅
対応付いた罫線特徴の識別番号
罫線特徴２
罫線特徴ｎ１
罫線特徴ｎ２
表特徴２

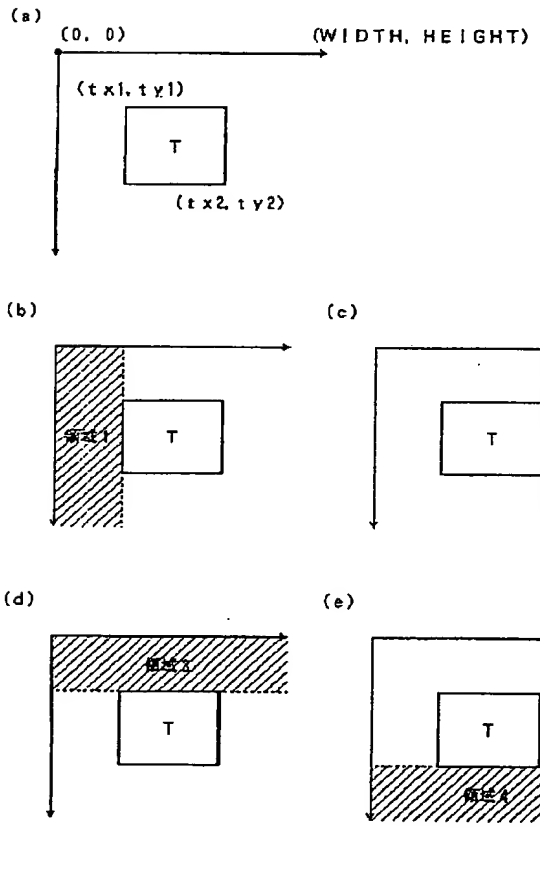
【図１７】



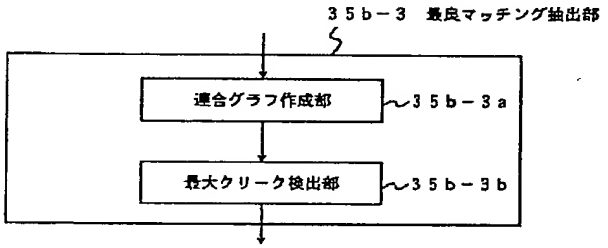
【図３１】



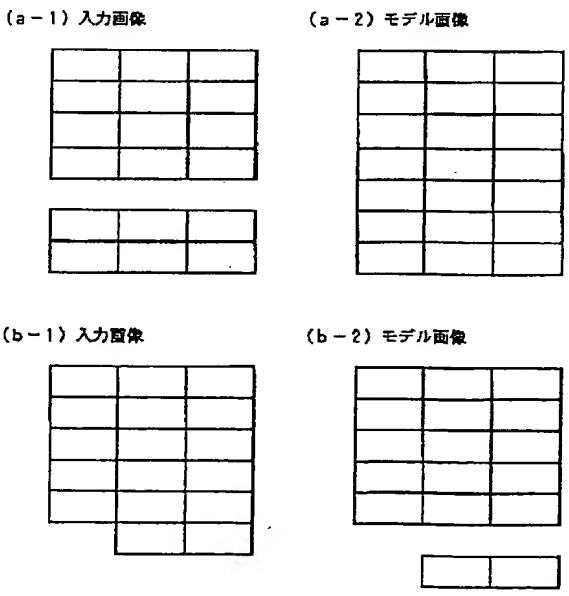
【図１８】



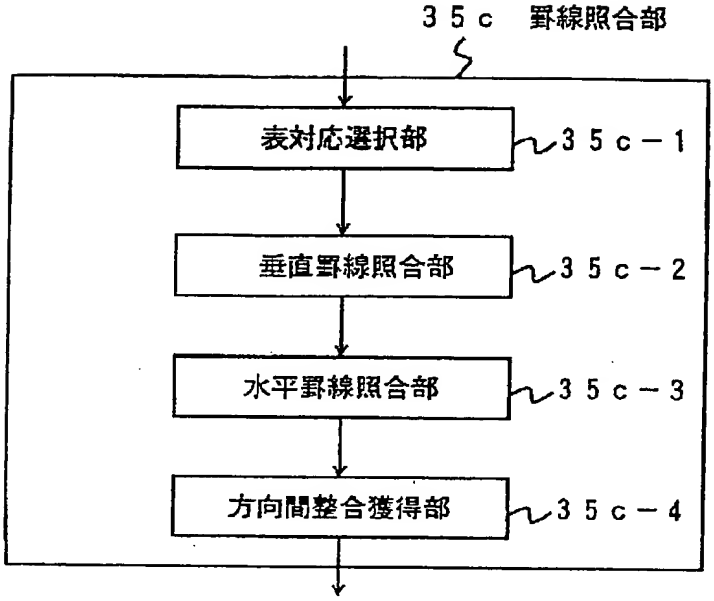
【図１９】



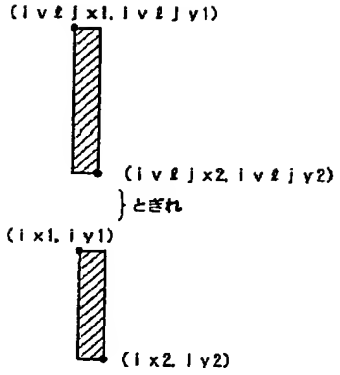
【図２１】



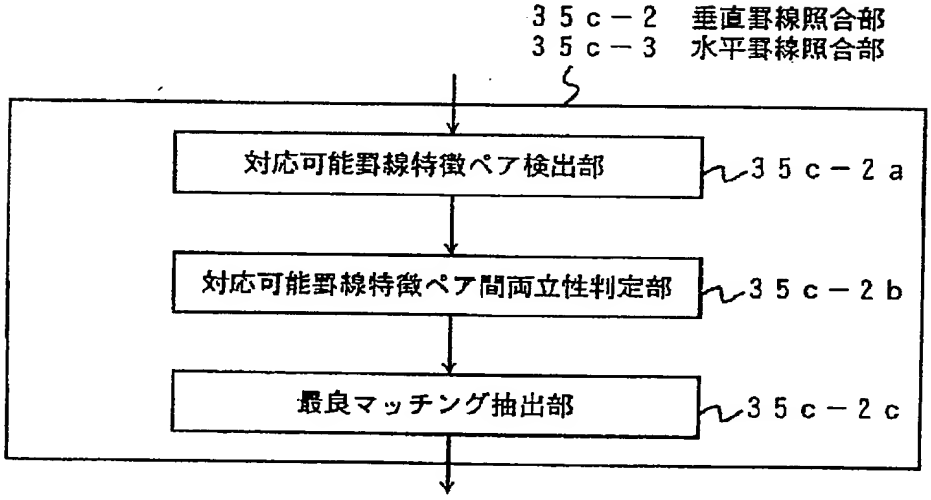
【図２２】



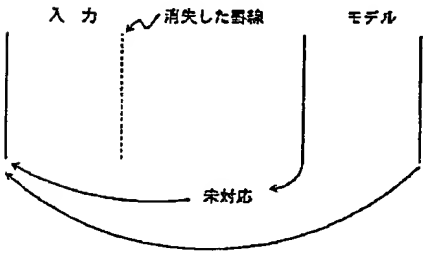
【図２４】



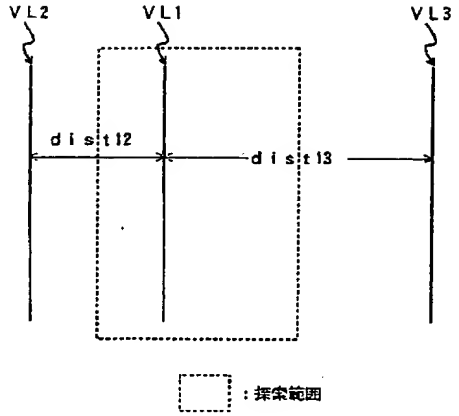
【図２３】



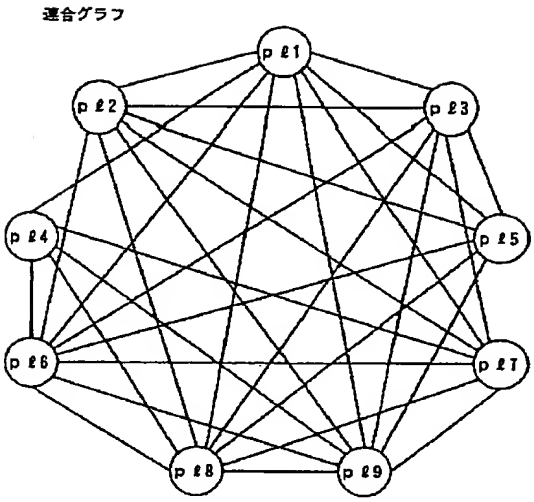
【図２９】



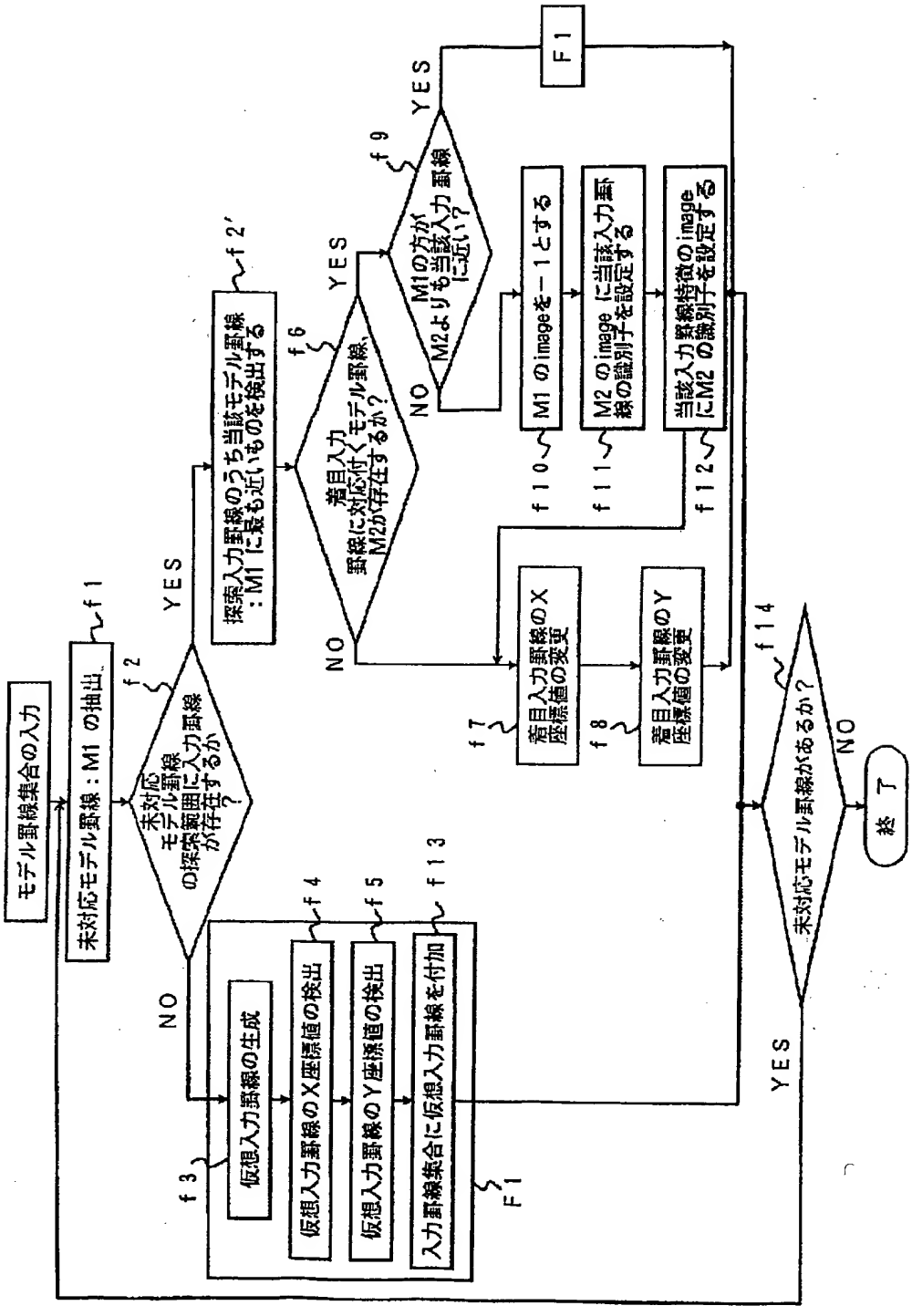
【図２７】


VL2
VL1
VL3
dist12
dist13
：探索範囲

【図２８】


連合グラフ
p ℓ1
p ℓ2
p ℓ3
p ℓ4
p ℓ5
p ℓ6
p ℓ7
p ℓ8
p ℓ9

【図30】


モデル罫線集合の入力
未対応モデル罫線：M1 の抽出
f1
f2
未対応
モデル罫線
の探索範囲に入力罫線
が存在するか？
YES
探索入力罫線のうち当該モデル罫線
：M1 に最も近いものを検出する
f2'
f6
着目入力
罫線に対応付くモデル罫線、
M2が存在するか？
YES
f9
M1の方が
M2よりも当該入力 罫線
に近い？
YES
F1
NO
M1 のimageを－1とする
f10
M2 のimage に当該入力罫
線の識別子を設定する
f11
当該入力罫線特徴のimage
にM2 の識別子を設定する
f12
NO
着目入力罫線のX
座標値の変更
f7
着目入力罫線のY
座標値の変更
f8
NO
仮想入力罫線の生成
f3
仮想入力罫線のX座標値の検出
f4
仮想入力罫線のY座標値の検出
f5
入力罫線集合に仮想入力罫線を付加
f13
F1
f14
未対応モデル罫線があるか？
YES
NO
終了

Part translation of JP

Published Japanese Patent Application No: H08-77294

(Paragraphs 0052-0057)

[0052] Fig.5 is a drawing of an example set of line segments the line-segment extracting unit 27a extracts from an input image. From the drawing of Fig.5, it becomes clear that the extracting unit 27a extracts, at the stage associated with this drawing, character-constituting short-length line segments in character areas of the input image. It is possible to eradicate these short line segments by processes as explained here. A candidate character rectangle extracting unit 27b is employed and configured to select patterns (hereafter called candidate character rectangles) each being formed by a set of mutually connected black pixels and judged to represent a character in advance of or following the line segment extraction process, or even in parallel to the line segment extraction process. The candidate character rectangle selection process comprises:

[0053] Step 1: Extracting rectangles each circumscribing a pattern formed by a mutually connected set of black pixels and determining the associated height ch and width cw for each of these rectangles.

[0054] Step 2: Determining the most frequently appearing values with respect to all these ch and cw values and adopting the most frequently appearing ch and cw values, respectively, as the average height value (CH) and average width value (CW) of characters contained in the input document.

[0055] Step 3: Determining patterns each formed by a mutually connected set of black pixels as those to represent the candidate character rectangles when the ch and cw values respectively associated with the patterns satisfy following formulas and extracting these patterns determined to represent the candidate character rectangles.

$$(CW - th) \leq cw \leq (CW + th)$$, and further $$(CH - th) \leq ch \leq (CH + th)$$

[0056] Fig.6 is a drawing to illustrate examples of candidate character rectangles (in which the character areas are respectively indicated by circumscribing rectangles) the candidate character rectangle extracting unit 27b extracts.

[0057] With respect to the above processes, it is possible to perform extraction of these candidate character rectangles in advance of performing extraction of line segments, and subject the part of the input image other than these candidate character rectangles to below-described filtering processes of both vertical and horizontal directions so that the line segment extracting process can produce a more reliable result even under situations in which the ruled lines are blurred and/or broken in places.

=End=